

CARGO CULT MATH

THE SOFTMAX DELUSION

The Architecture of Artificial Certainty

NABIL EL MAHYAOUI

The Softmax Delusion

The Softmax Delusion

The Architecture of Artificial Certainty

Nabil El Mahyaoui

A must-read ebook for engineers, philosophers, and risk managers

Copyright © 2026 Nabil El Mahyaoui.
All rights reserved.

Contents

1 The Procrustean Bed **1**

1.1 What the Model Actually Computes . . 2

1.2 The Confidence Illusion 4

1.3 The Temperature Proof 5

2 The Cross-Entropy Trap **7**

2.1 The Arrogance Gradient 8

2.2 The Label Smoothing Confession 9

2.3 The Next-Token Mirage 10

3 The Map Is Not the Territory 12

3.1 Three Theories, Zero Matches 12

3.2 Syntactic Probability 14

3.3 The Geometry of Delusion 15

4 The Flattening of Uncertainty 17

4.1 Aleatoric vs. Epistemic 18

4.2 The Broken Thermometer 19

4.3 Why Calibration Is Not Enough 20

5 The ROI of Certainty 22

5.1 Why Business Loves the Lie. 22

5.2 The Anthropomorphism Trap 24

5.3 The Audit Problem 25

6 The Hallucination Economy **27**

6.1 The Circular Fix 28

6.2 Toward Honest Machines 29

6.3 The Category Error of Our Time 30

References **32**

1

The Procrustean Bed

In Greek myth, Procrustes offered travellers a bed guaranteed to fit. Too tall? He cut the legs. Too short? He stretched them. The bed always fit. The guest rarely survived.

Softmax is the Procrustean bed of artificial intelligence.

Every model that outputs a "probability" passes its final computation through the same function. It takes a vector of raw numbers—

the model's unfiltered internal state—and forces it into the shape of a probability distribution. The output always sums to one. Always looks confident.

The question no one asks: *was the input ever a probability in the first place?*

1.1 What the Model Actually Computes

A neural network's final layer produces *logits*: one real number per possible output. These are the result of billions of matrix multiplications cascaded through hundreds of layers.

A logit is not a probability. It is not a confidence level. It is a *real number on an unbounded scale*, representing the degree to which the network's internal geometry "points toward" a particular output. Logits are geometric arti-

facts: projections of a high-dimensional state onto output directions. They have no inherent probabilistic meaning.

Then softmax arrives:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

Three things happen simultaneously. **Exponentiation** amplifies differences: a logit gap of 2 becomes a ratio of 7.4; a gap of 5 becomes 148. The exponential is a *magnifier of confidence*. **Normalisation** divides by the sum, forcing the output to sum to 1—the illusion of probability. **Suppression of silence**: there is no output for "I don't know." Every input produces a confident output.

The machine lacks the vocabulary of silence.

1.2 The Confidence Illusion

When a model is genuinely uncertain, the logits are noisy, roughly uniform. Softmax converts this noise into 33.2%, 33.1%, 33.7%. This is not the model saying "I'm confused." It is a *specific, positive assertion about a uniform distribution*. The difference is the difference between a witness saying "I didn't see anything" and one saying "I saw three people, each equally likely to be the suspect."

When the model has strong activation, softmax converts $[8.3, 1.2, 0.4]$ into $[99.2\%, 0.7\%, 0.1\%]$. But what does 99.2% measure? Not the probability of being correct. Not a calibrated belief. It measures the *relative magnitude of geometric projections in a vector space*.

When a model says it is 95% sure, it is not expressing a belief. It is reporting that the cosine similarity between the current context and one answer is much larger than for the alternatives. Mistaking geometry for belief is the fundamental error of the AI age.

1.3 The Temperature Proof

Every language model has a *temperature* parameter that scales logits before softmax: $\text{softmax}(z_i/T)$. At $T < 1$, the model becomes "more confident." At $T > 1$, "less confident."

An engineer can make the same model, same weights, same internal state, express 60% or 99% confidence on the same output by adjusting a single number.

If confidence were a genuine property of understanding, it could not be arbitrarily adjusted with a scalar multiplication. The fact that it can be is proof that softmax outputs are not beliefs. They are *formatting choices*.

We are building civilisation-scale decision systems on numbers whose magnitude is controlled by a knob.

2

The Cross-Entropy Trap

If softmax is the mechanism of the delusion, cross-entropy loss is its *incentive structure*—the mathematical definition of "success" that every modern neural network is trained to minimise. It contains a deeply underappreciated bias: it *rewards overconfidence*.

2.1 The Arrogance Gradient

During training, cross-entropy measures how "surprised" the model is by the correct answer:

$$\mathcal{L} = -\log p(\text{correct})$$

At 0.9 confidence, loss is 0.105. At 0.99, it is 0.01. At 0.999, it is 0.001. The loss always decreases as confidence rises. There is *no penalty for overconfidence*. The globally optimal solution is 100% confidence on every correct answer—a degenerate, infinitely overconfident distribution.

No real model achieves this, but the training dynamics *push in this direction*. The longer you train, the more overconfident the model becomes. Modern neural networks, straight out of training, are systematically miscalibrated:

they assign 95% confidence to predictions that are correct only 75% of the time.

> *We do not train models to be truthful. We train them to be arrogant.*

The industry fix is *temperature scaling*—post-hoc adjusting softmax temperature to improve calibration. Note the circularity: we train the model to be overconfident, then manually tune the amplitude of the lies until they are statistically acceptable.

2.2 The Label Smoothing Confession

There is a technique called *label smoothing*: instead of training on hard labels (100% correct, 0% everything else), you distribute a small probability mass to incorrect classes. It

improves generalisation. Reduces overconfidence. Produces better models.

Consider what this means: we get better results by *lying to the model about the truth*, because the model's notion of truth (100% confidence in one answer) was never truthful to begin with. Label smoothing is a confession: *we know cross-entropy produces false certainty, and we are manually injecting doubt to counteract it.*

2.3 The Next-Token Mirage

In language models, cross-entropy penalises the model for not predicting the *specific* next token that appeared in the training data. But language is not deterministic. After "the cat sat on the," the next word could be "mat," "roof," "chair," or a hundred other nouns.

The loss function sees only the continuation that appeared. Over millions of examples, this pressure forces the model to learn what is *frequent in the corpus*, not what is *true in reality*. A statement appearing ten thousand times will be assigned higher probability than one appearing once, regardless of accuracy.

This is why models hallucinate with high confidence. The hallucination is not a bug in the confidence mechanism. It *is* the confidence mechanism, functioning exactly as designed: assigning high probability to outputs that pattern-match the training data, whether or not they correspond to anything real.

3

The Map Is Not the Territory

We have built a map (softmax outputs) that satisfies all the formal properties of the territory (probability theory) while bearing no necessary relationship to it.

3.1 Three Theories, Zero Matches

Frequentist: probability is the long-run frequency of an event in repeated trials. A soft-

max output is not a frequency. The model is not reporting how often it would give the same answer. There is no ensemble, no repeated trial.

Bayesian: probability is a degree of belief, updated by evidence via Bayes' theorem. A softmax output is not a posterior. The model has no explicit prior, no likelihood function, performs no Bayesian updating. It performs gradient-trained pattern matching followed by a normalisation trick.

Propensity: probability is a physical property of a system. A softmax output is obviously not a propensity. There is no physical indeterminacy in a deterministic matrix multiplication.

Softmax outputs are compatible with none of these. What are they?

3.2 Syntactic Probability

They are **syntactic probabilities**: numbers satisfying the axioms of probability theory (non-negative, sum to one) without satisfying the *semantics* of any interpretation. They follow the grammar without the meaning.

"Colourless green ideas sleep furiously" is syntactically valid English and semantically void. Similarly, $[0.85, 0.10, 0.05]$ is syntactically valid probability and semantically void—unless we can establish a connection between those numbers and some property of the world.

Every product displaying a "confidence score," every system thresholding on a probability cutoff, every risk model feeding softmax outputs into a decision framework is conflating syntax with semantics.

We have the runways, the control tower, and the radar screens. We have the headphones and the coconut shells. We are performing all the rituals of probability. But the planes are not coming.

Cargo Cult Math. The form without the substance.

3.3 The Geometry of Delusion

The model's internal state is a vector in thousands of dimensions. The output layer projects this vector onto a set of directions—one per class. The logits are *dot products*: measures of geometric alignment, not of truth.

If the model's representation happens to be geometrically close to the direction labelled "cat," the softmax output for "cat" will be high. This

does not mean the input is a cat. It means the model's geometry arranged itself so that cat-like inputs land near the cat direction.

For in-distribution inputs, this often works. For out-of-distribution inputs, it can assign 98% confidence to something completely wrong, because the input's projection happens to align with a particular direction.

The system has no epistemic state. It has a *geometric configuration*, and we are projecting our psychological concept of certainty onto it.

4

The Flattening of Uncertainty

There are two fundamentally different kinds of not-knowing, and the conflation of them may be the most dangerous consequence of the Softmax Delusion.

4.1 Aleatoric vs. Epistemic

Aleatoric uncertainty: the world is random. A fair coin is 50/50. More data does not help. It is irreducible—a property of reality.

Epistemic uncertainty: you lack knowledge. Who won the 1904 Olympic marathon? The answer is definite, but if you do not know it, you are epistemically uncertain. More data helps. It is reducible—a property of the observer.

The distinction is essential. If a medical model is uncertain because the disease is genuinely ambiguous (aleatoric), order more tests. If it is uncertain because it has never seen this case (epistemic), flag it for human review. Completely different responses to completely different problems.

Softmax destroys this distinction. A single number: 0.6. Why 0.6? Genuine ambiguity? Insufficient training data? Adversarial input? The number cannot say. It *flattens* the rich landscape of uncertainty into a scalar. All information about *why* the model is uncertain—the information essential for correct decisions—is discarded.

4.2 The Broken Thermometer

Imagine a thermometer that always gives a reading. Even when it is not immersed in anything. Even when its mercury is cracked. It never reads "ERROR." It always displays a plausible number.

Now build a climate model on its readings. On good days, it is accurate. But you cannot distin-

guish good readings from bad ones, because the instrument never signals malfunction.

This is softmax. Every input—no matter how far from training distribution, no matter how adversarial, no matter how nonsensical—produces a crisp distribution summing to exactly one.

4.3 Why Calibration Is Not Enough

The standard fix is "calibration": post-hoc adjusting outputs so that, on average, 80% confidence means 80% accuracy. Three problems:

Population, not instance. A model can be perfectly calibrated on average while being catastrophically wrong on individual predictions. For any specific input, the number tells you nothing about whether *this* prediction is right.

Distribution-dependent. Calibration holds for the distribution it was measured on. Out-of-distribution—exactly where uncertainty matters most—the guarantees evaporate.

Still flat. A calibrated model still produces one number. It still cannot distinguish aleatoric from epistemic uncertainty. The flattening persists.

Calibration is a Band-Aid on a structural wound. It improves the thermometer's average accuracy without giving it the ability to signal when it is broken.

5

The ROI of Certainty

The Softmax Delusion is not merely a technical curiosity. It is a *market structure*.

5.1 Why Business Loves the Lie

Corporations are certainty-seeking organisms. A consultant who says "I think so, but I'm not sure" gets replaced. A consultant who says

"there is an 83% probability of success" gets a contract extension.

The number does not need to be correct. It needs to be *specific*. Specificity provides cover. When the decision goes wrong, the decision-maker points to the number: "The model said 83%." The number transforms institutional responsibility into mathematical fate.

A credit model says "72% likelihood of default"—the loan officer decides without thinking. A hiring algorithm says "89% match"—the manager shortlists without reading. A triage system says "94% benign"—the doctor moves on. In every case, the number is treated as a measurement of reality. In every case, it is a geometric projection normalised to sum to one.

The return on investment of certainty is not measured in accuracy. It is measured in *decision velocity*. The fact that some fraction of those decisions are wrong is acceptable, because the cost of being wrong is distributed and delayed, while the benefit of being fast is immediate and concentrated.

5.2 The Anthropomorphism Trap

When a model outputs 0.95, we do not experience it as "the ratio of exponential projections in a vector space." We experience it as "the system is very confident." We attribute a mental state—confidence—to a machine that possesses no mental states.

This is not a failure of education. It is a feature of human cognition. We are social primates, evolved to infer mental states from be-

havioural signals. When a system produces outputs *isomorphic to confidence*, our social inference machinery activates automatically.

The AI industry knows this and exploits it. Confidence scores are displayed as progress bars, colour-coded green for high and red for low. Nowhere in the interface does it say: "This number is a normalised exponential of a dot product and has no guaranteed semantic content."

5.3 The Audit Problem

The most consequential arena is regulated industry: finance, healthcare, criminal justice. Consider algorithmic trading: billions move on softmax distributions over market moves. Consider sentencing: a defendant's liberty may depend on a COMPAS score that is, in

the strictest technical sense, *not a probability of anything*. Consider diagnostics: a "93% benign" that discourages a biopsy.

The audit problem is that *no one in the decision chain understands what the number means*. The engineer knows it is a softmax output. The product manager calls it a "confidence score." The end user treats it as a probability. The regulator evaluates it as a calibrated risk estimate. Each layer of abstraction adds a layer of interpretation, and each moves further from mathematical reality.

6

The Hallucination Economy

The AI industry's most pressing problem is hallucination: models that produce fluent, confident, and entirely fabricated outputs. Billions of dollars target the symptom. Almost every proposed fix relies on the same tool that causes the problem: the model's own confidence scores.

If your trading bot or diagnostic tool relies on the raw confidence score of an

LLM, you are not doing data science.
You are doing astrology with vectors.

6.1 The Circular Fix

The dominant approach to hallucination detection is confidence thresholding: if the softmax probability falls below a threshold, flag the output as unreliable.

This assumes softmax probabilities correlate with correctness. For in-distribution inputs, the correlation is imperfect but real. For out-of-distribution inputs—exactly where hallucination matters—the correlation can *invert*: the model hallucinates with high confidence precisely because it has no training signal to indicate unfamiliar territory.

We are using the model's self-assessment to detect confident hallucinations. This is asking the suspect to evaluate the credibility of their own alibi.

6.2 Toward Honest Machines

What would it take to escape the delusion? Not incremental fixes. *Architectural* changes.

Separate answer from confidence. Train a dedicated uncertainty module that evaluates the model's internal state independently—optimised not for cross-entropy but for *calibrated error prediction*.

Ensemble disagreement. Run the input through multiple models and measure disagreement. High disagreement is a model-independent uncertainty signal: "different

reasonable models disagree" is a genuine epistemic statement.

Conformal prediction. Produce a *set* of plausible answers with a statistical guarantee: "the correct answer is in this set with 95% probability." A set of size 1 means confidence. A set of size 50 means ignorance. Uncertainty expressed as *coverage*, not confidence.

Abstention. Give the model the ability to output "I don't know" as a first-class response with its own reward signal. A model that can abstain is a model that can be silent—the capacity softmax most thoroughly destroys.

6.3 The Category Error of Our Time

We have built a multi-trillion-dollar industry on a function that forces any vector into the

shape of a probability distribution. We have treated the shape as the substance. Mistaken the syntax of certainty for the semantics of knowledge.

Breaking out requires a conceptual shift: a number summing to one is not, by that fact alone, a probability. Confidence is not competence. The map is not the territory.

The delusion ends when we stop asking "how confident is the model?" and start asking "what are the model's reasons, and are they good ones?"

That question cannot be answered with a number between zero and one. It can only be answered with thought, scrutiny, and the irreducibly human capacity for doubt.

References & Further Reading

Guo, C. et al. (2017). "On Calibration of Modern Neural Networks." *ICML*.
The landmark paper demonstrating that modern neural networks are systematically overconfident.

Feynman, R. (1974). "Cargo Cult Science." Caltech commencement address.
On the difference between the appearance and the substance of scientific reasoning.

Gal, Y. and Ghahramani, Z. (2016). "Dropout as a Bayesian Approximation." *ICML*.

A practical method for estimating epistemic uncertainty in neural networks.

Szegedy, C. et al. (2016). "Rethinking the Inception Architecture." *CVPR.*
The paper that introduced label smoothing and inadvertently revealed the overconfidence problem.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.
The theoretical foundation of conformal prediction—probability without the delusion.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
On the cognitive biases that make humans vulnerable to false confidence signals.

Taleb, N. N. (2007). *The Black Swan*. Random House.

On the catastrophic consequences of confusing measured risk with actual uncertainty.

About the Author

Nabil El Mahyaoui

https://nabil-el-mahyaoui.com/

https://www.linkedin.com/in/elmahyaoui