A FORMALISM IN SEVERAL FRAMINGS? EXPLICATING $-\log p(x)$ IN MACHINE LEARNING

Jan KOSTANJEVEC

Independent researcher

Keywords: explication, negative log-probability, conceptual variation, computational philosophy of science, arXiv

1 INTRODUCTION

Carnap, in *Logical Foundations of Probability* (Carnap, 1950/1971), and subsequent philosophers of probability were faced with the following conceptual constellation: a single formalism (the probability calculus) sustaining multiple, mutually irreducible interpretations (frequentist, subjective, logical, propensity). A similar pattern can be observed in machine learning literature, but without the accompanying philosophical discussions. The negative log-probability of data under a model ($-\log p(x)$) appears as summed, averaged, or exponentiated across contemporary practice under at least five distinct names: *log-likelihood* (statistics, after Fisher, 1922), *cross-entropy loss* (vision and general ML), *negative log-likelihood* or *NLL* (probabilistic ML and reinforcement learning), *perplexity* (speech and natural language processing, after Jelinek et al., 1977), and *training loss* (engineering-oriented deep learning). The formal object is the same up to sign and monotone transformations and the names seem to index different sub-communities. The present study builds an instrument for detecting whether and how the framings vary conceptually.

2 THEORETICAL FRAMEWORK

This project is positioned within computational philosophy of science (Malaterre et al., 2021; Pence & Ramsey, 2018) and draws on the recent revival of Carnapian explication (Brun, 2016; Carus, 2007; Dutilh Novaes & Reck, 2017; Justus, 2012) with two important emphases.

First, Carnap's "natural concept" starting point is loosened: following Justus (2012) and Shepherd and Justus (2015), the explicandum is understood as possibly already a formalised scientific concept whose identity is fixed, e.g. by the epistemic goal it addresses (Brigandt, 2010).

Second, Carnap's "explicatum" is explicitly analysed in two parts as in Carnap (1950/1971), so the resulting typology has three slots: explicandum (C), formalism (F), and interpretation (I). Each has a cardinality 0, 1, or N, subject to the current state of use. Different combinations can be realised, and this study attempts to determine which are ruled out on logical and empirical grounds.

The typology spanning the C/F/I axis is distinct from Brigandt (2010)'s own triadic decomposition of conceptual content (reference, inferential role, and epistemic goal), which sits inside the explicatum. The most consequential cell for this paper is $(1, 1, N)$, i.e. one explicandum, one formalism, contending interpretations. This is also the cell Carnap analysed in his studies of the probability calculus. The study presents an instrument designed to test whether $-\log p(x)$ occupies (1,1,N) or (N,1,N). However, the sample in the present study is too small to fully run this test.

Additionally, the "paradox of adequate formalization" identified by Dutilh Novaes and Reck (2017), i.e. the tension between Carnap's similarity and fruitfulness criteria, is less corrosive from our perspective. If one formalism sustains multiple interpretations each fruitful for a different sub-community, then no single explicatum needs to be adequate for all explicanda.

3 RELATED WORK

At least three areas of prior work are related to this project. Simons (2024) uses BERT to diachronically analyse physics terms and track shifts in the use of scientific terms. Contributions from computational philosophy of science, such as Malaterre et al. (2021)'s topic-modelling analyses of philosophy-of-science journals and the methodological manifesto of Lean et al. (2023) are examples of using corpus methods in philosophy of science.

Scharpf et al. (2023)'s *Formula Concept Discovery* has shown that formal expressions can be objects of analysis in scientometrics. Perhaps the closest neighbour is Oyama et al. (2025)'s "Mapping 1,000+ Language Models via the Log-Likelihood Vector," which deals with the cross-entropy / NLL / perplexity identity operationally when mapping their model but doesn't attempt a conceptual analysis.

Teufel and Moens (2002)'s Argumentative Zoning and Liakata et al. (2012)'s CoreSC established sentence annotation of categories that aren't apparent on the surface text. The present study follows this path. The formal equivalence of different framings of $-\log p(x)$ is explicitly treated as a technical identity in canonical textbooks (Goodfellow et al., 2016; Jurafsky & Martin, 2026; Murphy, 2025) and often acknowledged

operationally in practice, but never as a target of conceptual analysis. Our project is therefore an interpretive analysis of a single formal expression, reading its specific framings against a Carnapian account of conceptual variation.

4 METHOD

The corpus is drawn from arXiv `.tex` sources rather than PDFs since extracting data from PDFs would introduce additional noise. However, extraction from `.tex` surfaced issues that belong to reading the source rather than the rendered text, for example instead of layout artifacts, we might get instances of target strings in comments, filenames and macro definitions. The `.tex` files were limited to the following arXiv categories: cs.LG, cs.CL, stat.ML. From each category, the last 100 papers submitted in June 2024 were selected. Since paper volume varies across categories, the number of days covered for each category differs. The submission-announcement gap is the reason some papers have an id for July 2024, since ids are awarded on announcement not submission. This resulted in 276 unique papers, 23 of which had overlapping categories (which were assigned deterministically to one category).

In this sample we matched loosely two types of string patterns: math notation (`\log p`, `-\sum...\log`, `\mathrm{PP}`) and natural-language terms ("perplexity", "log-likelihood", "negative log-likelihood", "cross-entropy loss", "training loss").

This hybrid design was chosen since a name can be used instead of a formalism (a named quantity can appear with no formula on the page) and the aim was to capture conceptual differences both with respect to the formalism and namings (a single name might be imbued with a different conceptual frame despite naming the same formalism). The matching resulted in 914 passages, almost evenly divided in half between the two types.

A codebook was constructed, partly inspired by the Argumentative Zoning and CoreSC approaches to sentence/paragraph level annotation of interpretively loaded categories (i.e. those that can't be classified by surface text alone). However, agreement statistics aren't reported because there was only one annotator (see §Limitations). The codebook required each occurrence to be screened for whether it is a bona fide usage in the first place and then to classify it into one of four theory-derived framings (FR1–4) or a miscellaneous category for underdetermined passages (FR5). The four theory-derived framings are FR1: information-theoretic (bits, code lengths), FR2: likelihoodist (estimations, model fit), FR3: optimisation-centric (e.g. a quantity that's being minimised) and FR4: evaluative (benchmarks and scores). The framings

were deliberately neutral with respect to the theory of explication (neither explicandum nor interpretation roles are assumed at this stage).

For each passage the annotator also answered two probe questions (1) the epistemic goal (Brigandt, 2010) of the use of target in a given passage and (2) the mapping, which asks what type of quantity x is in $-\log p(x)$, in order to track conceptual variation between and within framings. These are compared across passages along with a zeugma probe, testing if different usages can sensibly be conjoined in a single predicate. The present sample is too small to give systematic patterns (see Limitations), so we present only indicative examples in the results.

After several version-controlled iterations the codebook was frozen after which the author carried out the annotation of 18 passages, where the target occurrence was marked, since multiple possible targets can be present within a given passage. They were drawn in random order from a seeded random subset of 20 passages per category of the total cleaned set of passages (excluding duplicates and bare symbols) consisting of 465 passages. During annotation the arXiv categories and the paper's id were hidden, the section heading of the passage was shown.

5 RESULTS

Across 276 papers 97 contained a match totaling 1077 targets matched, 163 were comments, and from the rest, 914, the corresponding passages were extracted. Roughly half (472) were based on notation patterns, while the name patterns accounted for 442 matches. 588 passages passed automatic screening which attempted to rule out the following types of noise: bare symbol use (target appearing without interpretive terms in the preceding or following 400 characters) was found in 211 cases (164 of these originated from "latex:log p", the loosest notational pattern); 92 appeared in captions or tables and 61 in "display math" environments, all without ambient prose. Some passages failed on multiple automatic screens. Another source of noise was found in filenames and paths, which were not automatically screened.

We can see that notational patterns account for 472 out of 914 overall matches, but resulted in 240 out of 326 screened out passages. This indicates that notational patterns are more likely to return noise, symbols in derivations rather than usages that would afford an interpretation of the frame.

Of 18 annotated passages 12 were codeable, 6 were skipped (5 were uses in derivation or similar, 1 was a filename occurrence). From the 12 codeable passages 5 were coded as FR3, 4 as FR4, and 1 each as FR1, FR2, and FR5. All four theory-derived framings

were attested. 1 passage was coded as FR5 because it was deemed underdetermined. No evidence was found for additional framings beyond the four theory-derived ones, although the sample is too small to claim exhaustiveness of frames. The sample was large enough to surface multiple contrasts, which the probes were meant to detect. We present two illustrative contrasts, one across framings and one within a framing. The target occurrence is shown in bold.

For the first example consider the following passages, the first we read as FR3:

> Another possibility that would be relevant where obtaining a high level of equivariance is more important than the sheer accuracy of the energy and force predictions, is to modify the **training loss** to explicitly penalise symmetry breaking

and the second we read as FR2:

> [...] iterate over each sentence, masking a word at a time (except for the words that identify a social group), and accumulate the **log-likelihood**s of the masks in a sum for comparison.

Here we see the same formal object operating in two different framings, with two different epistemic goals. In the first, the goal is to shape a model to have a desired property, in the second, the goal is to measure an (ostensibly social!) property of a model. The first is a malleable component of a training process; the second is a measuring instrument for a property of elements within the trained model. The difference here is not just what the quantity is called, but what it's for.

For the second example contrast two passages from the same frame (FR4):

> [...] evaluate the complexity and readability of the text, with pages having a **perplexity** score over 1,000 being discarded

> Considering any time step of the "no pruning" baseline, we see that perplexities decrease with longer files, up to the context length (black dotted line), as we'd expect since there's more "useful" context enabling the model to lower **perplexity**. Beyond the context length, perplexities increase as expected (since all documents are being cut somewhere in the middle). Surprisingly, perplexities again decrease on the *longest documents*, seeming to indicate that being split isn't affecting these documents. This is in line with our qualitative observations of repetitiveness [...].

Both targets have the same name, both passages are in the evaluative framing, and yet they have different epistemic goals. In the first it is used as a proxy for the limit of useful text, in the second it is used to evaluate pruning strategies. This shows that the framings themselves could be coarser than the conceptual material they subsume. If, however, it turns out that these framings are enough for conceptual individuation the probes might not track essential aspects of concepts.

The mapping probe showed that it's rare to infer the content of p from the extracted passages, with 9/12 fields being explicitly coded as "not stated". If this generalises, it seems that the mapping probe isn't useful at the scale of the fragments we worked with.

6 CONTRIBUTIONS

Four contributions are presented. First, theoretical, an extension of Carnap's theory of explication via a triadic typology (C, F, I, with cardinality 0, 1, or N at each), distinguished from Brigandt (2010)'s intra-explicatum triad and offered as a small reusable analytical tool for further work on the conceptual organisation of formal sciences.

Second, methodological, a versioned and frozen annotation instrument for interpretively loaded categories spanning a formal expression is released. It extends the Argumentative Zoning and CoreSC traditions.

Third, empirical, naive matching yields mostly non-usages which were classified (bare symbols, comments in source, filenames and captions). Notation patterns are the main source of this kind of noise.

Fourth, conceptual, the probes are an attempt at operationalising an evidential frame for conceptual divergence and the illustrative contrasts demonstrate distinguishable conceptual traits. If it turns out that framings diverge systematically then comparisons that cross framings might be comparing quantities imbued with incompatible epistemic goals.

Additionally, the study surfaced theoretical issues from application and hopes to contribute better tools to track conceptual variation in relation to formal systems within philosophy of science. This instrument should be applicable to a wide range of formal expressions that may or may not differ in their namings.

The source code, the codebook and sample manifests will be published under a FOSS license. The raw arXiv sources will not be redistributed, but an included script will reproduce the corpus.

7 LIMITATIONS

The study has several limitations. The first concerns formalism identity. The target expressions are equal only up to sign, monotone transformation, and aggregation. Perplexity is exponentiated, surprisal is pointwise, cross-entropy equals mean NLL

only when the reference distribution is the empirical one, and "training loss" names a role rather than a formula. The internal structure and individuation of F remain open for future work.

Annotation was done by only one annotator, despite the study being designed for more, moreover the annotator was a philosopher without expert familiarity with ML practice. Scope was reduced to one month so no diachronic, prevalence or community specific claims are made. The small n of completed annotations (18) supports existence and illustration only. The annotations were not recoded for intra-rater stability, thus no agreement statistics are reported.

The sampling mechanism gives a different span of article submission dates because of volume differences between categories. Thus, they cannot be matched on submission window.

While some false positives were caught before drawing the sample, some recall problems were missed: latex macro definitions escape the pattern matching approach and more names could have been included (e.g. "surprisal" and mere "likelihood").

Additionally, the keyword set was drawn from the framings and can exclude prose passages with unexpected vocabulary as "bare symbol" occurrences effectively screen out a portion of the unexpected framings.

8 DISCUSSION

Measuring bits, minimising a loss, deciding on a perplexity score and evaluating a model fit rely on doing a formally similar thing. Whether they are doing the same thing conceptually is a question echoing the one Carnap and other philosophers were asking when interpreting the probability calculus. This project takes that question seriously and treats it as inspiration for empirical studies.

REFERENCES

Brigandt, I. (2010). The epistemic goal of a concept: Accounting for the rationality of semantic change and variation. *Synthese*, *177*(1), 19–40. https://doi.org/10.1007/s11229-009-9623-8

Brun, G. (2016). Explication as a Method of Conceptual Re-engineering. *Erkenntnis*, *81*(6), 1211–1241. https://doi.org/10.1007/s10670-015-9791-5

Carnap, R. (1971). *Logical foundations of probability*. Univ. of Chicago Press. (Original work published 1950)

Carus, A. W. (2007). *Carnap and twentieth-century thought: Explication as Enlightenment*. Cambridge University Press.

Dutilh Novaes, C., & Reck, E. (2017). Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. *Synthese*, *194*(1), 195–215. https://doi.org/10.1007/s11229-015-0816-z

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594-604), 309–368. https://doi.org/10.1098/rsta.1922.0009

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT press.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, *62*(S1), S63–S63. https://doi.org/10.1121/1.2016299

Jurafsky, D., & Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd, draft).

Justus, J. (2012). Carnap on concept determination: Methodology for philosophy of science. *European Journal for Philosophy of Science*, *2*(2), 161–179. https://doi.org/10.1007/s13194-011-0027-5

Lean, O. M., Rivelli, L., & Pence, C. H. (2023). Digital Literature Analysis for Empirical Philosophy of Science. *The British Journal for the Philosophy of Science*, *74*(4), 875–898. https://doi.org/10.1086/715049

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, *28*(7), 991–1000. https://doi.org/10.1093/bioinformatics/bts071

Malaterre, C., Lareau, F., Pulizzotto, D., & St-Onge, J. (2021). Eight journals over eight decades: A computational topic-modeling approach to contemporary philosophy of science. *Synthese*, *199*(1-2), 2883–2923. https://doi.org/10.1007/s11229-020-02915-6

Murphy, K. P. (2025). *Probabilistic machine learning: An introduction* (Third printing). The MIT Press.

Oyama, M., Yamagiwa, H., Takase, Y., & Shimodaira, H. (2025). Mapping 1,000+ Language Models via the Log-Likelihood Vector. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 32983–33038. https://doi.org/10.18653/v1/2025.acl-long.1584

Pence, C. H., & Ramsey, G. (2018). How to Do Digital Philosophy of Science. *Philosophy of Science*, *85*(5), 930–941. https://doi.org/10.1086/699697

Scharpf, P., Schubotz, M., Cohl, H. S., Breitinger, C., & Gipp, B. (2023). Discovery and recognition of formula concepts using machine learning. *Scientometrics*, *128*(9), 4971–5025. https://doi.org/10.1007/s11192-023-04667-9

Shepherd, J., & Justus, J. (2015). X-Phi and Carnapian Explication. *Erkenntnis*, *80*(2), 381–402. https://doi.org/10.1007/s10670-014-9648-3

Simons, A. (2024, November). Astro-HEP-BERT: A bidirectional language model for studying the meanings of concepts in astrophysics and high energy physics. https://doi.org/10.48550/arXiv.2411.14877

Teufel, S., & Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, *28*(4), 409–445. https://doi.org/10.1162/089120102762671936