Commitment Before Competence in Neural Network Development

Philipp Moritz Stecher[1], Sandro Radovanović[2], and Reinhard Kahle[1,3]

[1] Carl-Friedrich-von-Weizäcker Zentrum, University of Tübingen, Germany
philipp.stecher@fsci.uni-tuebingen.de
[2] Faculty of Organizational Sciences, University of Belgrade, Serbia
[3] CMA and Departamento de Matemática, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Portugal

Abstract. Neural networks commit before they are competent. Early in training, while performance is still poor, a network settles on internal structure that later learning builds on, so by the time behaviour reveals which solution it chose, revising that choice means relearning whatever depends on it. Our two companion studies point in this direction, since blocking early learning of broad categories in vision delayed finer learning by the equivalent of 2.5 passes through the training data beyond a matched control, and activity from the first 1% of language-model training forecast the structural roles that internal components later occupied. We argue that mechanistic interpretability can read commitments in the gap before competence, and we propose measuring that gap and testing whether acting within it spends less training on correction and more on new abilities.

Keywords: neural network development · training dynamics · developmental commitment · interpretability · cognitive development
Position paper.

1 Introduction

Between six and twelve months of age, infants become better at hearing the sound contrasts of their native language and worse at hearing contrasts that it does not use [1]. At that age most have yet to say a first word. Kuhl and colleagues explain the pattern as neural commitment to the sounds of the native language. Hearing settles on one language before the child can speak it, and on their account this settlement supports the learning that follows while making the sounds of other languages harder to learn.

We argue that artificial neural networks develop in the same order. **Our position is that commitment comes before competence.** A network settles how it will solve a task before it can solve it, and its performance reports that choice only after later learning already depends on it. By *commitment* we mean settling on internal structure that later learning builds on. By *competence* we mean reliable performance on the task. The position is a claim about their order.

That order is easy to miss. Training is usually followed through error curves and benchmark scores, which track competence, while interpretability is mostly applied to finished models. If structure and performance arrived together, nothing would be lost. If commitment comes first, performance is a late report on a choice that is already being built upon.

Two of our own studies led us to this view. Consider a network learning to tell roses from tulips, which are both flowers. In our vision study, responses shared across such broad categories grew fastest early in training, and when we held that broad structure back for the first ten passes through the data, the finer distinctions were learned late [2]. In our language study, measurements from the first 1% of training forecast which internal components would later hold structural roles [3]. Neither study was designed to test the position. Together they are why we hold it.

The position makes three claims, and each can fail. The first is a claim of order, that internal measurements forecast a network's eventual solution before its performance does. The second is a claim of dependence, that changing the early structure changes what is learned afterwards. The third is a claim of rising cost, that the same correction takes more training the later it begins. To test them we define two quantities. The *commitment interval* is the lead of internal measurements over behaviour, and the *revision cost* is the extra training that a correction needs. If the interval is positive while the cost is still low, there is a window in which a commitment is visible and still cheap to change. Mechanistic interpretability, which studies the internal workings of networks, is the instrument for finding that window.

Networks permit an experiment that infants do not. We can save a learner in mid-development, copy it, change one thing in one copy, and let both continue. The tests we propose rest on this.

2 Two Studies That Point to the Position

Both studies follow a network's internal activity through training. Both use sparse autoencoders (SAEs), which approximate that activity as a combination of a few active components called *features*. We saved each network at intervals during training, giving a series of *checkpoints*, and tracked a feature from one checkpoint to the next by comparing its responses to the same inputs.

2.1 In vision, finer learning builds on early broad structure

We trained image classifiers on CIFAR-100, in which a hundred fine classes such as roses and tulips fall into twenty broad categories such as flowers [2]. Training used the fine labels alone, so any broad organisation had to form unprompted. We tracked three architectures on CIFAR-100, with ResNet-18 also tested on Tiny ImageNet. Across nine CIFAR-100 runs, responses shared within a broad category developed especially rapidly early in training, and responses separating the fine classes became more prominent later. The two kinds of learning overlapped, with the balance shifting over time.

We then tested whether the finer learning depends on the broad structure, using ResNet-18 on CIFAR-100. For the first ten epochs, or passes through the data, we blocked the part of each update to the network's weights that lay along directions separating the average responses of the broad categories. The network still processed images normally. A control blocked equally many directions elsewhere and matched the amount of blocked update as closely as possible, and a restricted comparison in later layers matched both exactly. Across five runs, blocking the broad directions delayed fine-category learning by a further 2.5 epoch-equivalents compared with the control (95% interval 1.4–3.5). We measured the delay by finding when ordinary training first reached the same validation accuracy.

The resulting features were organised differently too. Under broad-category blocking, 17% of class-selective features responded preferentially to other classes within their own broad category, against 34% under control blocking. In the reverse direction, blocking fine-category signals outside the broad structure largely spared broad learning. The study points to dependence. Competence at fine distinctions was built on broad structure that formed early, and withholding that structure changed both the pace and the organisation of what followed.

2.2 In language, later roles are foreshadowed early

We followed four language models, Pythia-160M, -410M and -1B and OLMo-2-1B, as they were trained to predict the next token of text [3]. In a trained model, some features sit above many narrower ones. We linked a narrow feature to a broader one when the broader feature covered most of its responses and their overlap exceeded chance, and we called persistent features with many narrower features beneath them *carriers*. Carriers were defined independently of the removal test below.

Six statistics of a feature's early activity predicted whether it would become a carrier, with AUC between 0.69 and 0.83 across the models. AUC is a ranking score on which 0.5 is chance and 1 is perfect. In Pythia the early measurement came at step 1,000 of 143,000. Early activity anticipated a feature's later role more clearly than whether it would survive. The roles forecast in this way matter to the finished model. Removing carriers raised next-token loss, the model's prediction error, 1.71–1.85 times as much as removing control features matched for layer, number, activation frequency and size of the removed signal. The study points to order. Which features will hold a structural role is partly readable within the first 1% of training.

2.3 What the studies leave open

Neither study tests the position, and their limits show where the tests must go. In vision, blocking directions estimated from random class groupings, which overlapped the broad-category structure, produced similar delays, so the effect may owe more to shared early structure than to meaningful categories. Whether

the effects persist after the block ends is untested, and a head start from broad-category pretraining cost twenty additional epochs, so no saving at equal total computation is established. In language, we evaluated forecasts on held-out features within the same models, our controls for tracking reliability were inconclusive for Pythia-1B, and removal can push internal activity outside its usual range. Whether early scores predict the later impact of removal remained inconclusive, and we have yet to change early features and continue training. The two settings also define broad structure differently, so a shared pattern remains a hypothesis. Finally, neither study compares internal measurements with a forecast from behaviour alone. That comparison is the first test we propose.

3 Why Commitment Should Come First

A commitment, in our sense, is a learned organisation of internal activity that later learning comes to depend on. It need not be permanent. Further training can strengthen, revise or replace it, and the features that carry it may change while its influence persists. What makes it a commitment is that revising it has a price, paid in relearning whatever was built on it.

Three observations make this the order to expect, one for each claim.

Easy regularities are learned early. Networks pick up simple statistics of their inputs before complex ones [4] and broad distinctions before fine ones [5]. Structure for these regularities is therefore in place while harder tasks remain unsolved, which supports the claim of order. On a task where competence arrives suddenly and late, measures of internal structure have been seen to improve well beforehand [6].

Learning is cumulative. Each training step adjusts the network that already exists, so whatever forms early is the material from which later abilities are made. This supports the claim of dependence.

Early states are hard to leave. A deficit imposed early in training can do lasting damage after it is removed [7,8]. Early in training a network becomes stable, in the sense that runs differing only in the order of their data end in the same connected region of solutions [9]. A network that continues from earlier training can generalise worse than one trained afresh [10], and long training on a sequence of tasks can erode the ability to learn new ones [11]. These results support the claim of rising cost.

Each of these lines of work holds part of the picture. Critical periods establish that timing matters, shortcut learning describes the limiting strategies on which finished networks turn out to rely [12], and loss of plasticity describes an end state in which further learning has become hard. Methods for tracking internal structure through training now exist [13,14]. The position connects these. It names the learned structure through which early training acts on later learning, claims that this structure can be read before behaviour shows it, and supplies quantities by which the claim can be checked.

A commitment can be a foundation or a trap. Grouping roses with tulips is a foundation, since finer distinctions are built on it. A classifier that tells cows

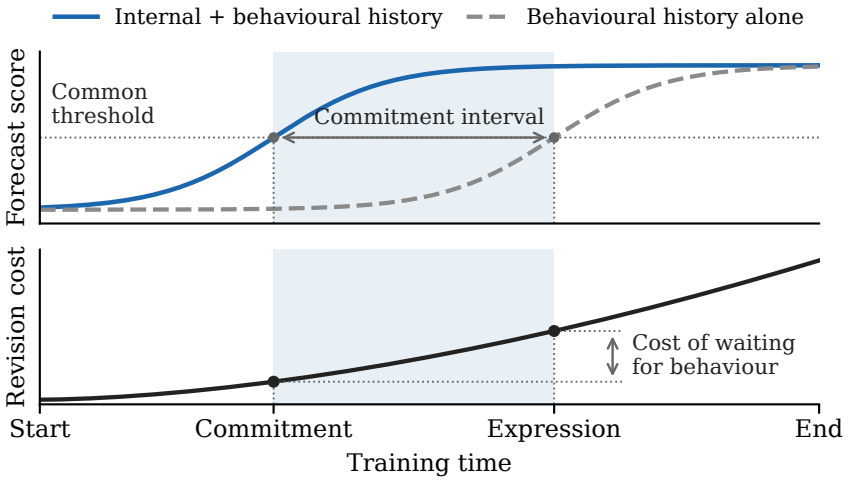

Fig. 1. Schematic of the two proposed quantities on a shared time axis. Top. Two forecasts of the same final outcome, with the commitment interval between their crossings of a common threshold. Bottom. The extra training needed to revise a commitment, by the time at which correction begins. The position predicts a positive interval and a cost that rises across it. No empirical results are plotted.

from camels by pasture and sand has built on a trap, and a cow on a beach exposes it [15]. The order is the same in both cases. What differs is whether we want to keep what was built first, so the practical question is the same as well. Can we see which kind is forming while it is still cheap to change?

4 Two Quantities That Make the Position Testable

The *commitment interval* measures how far behaviour lags behind internal structure. Take a property of the finished network, such as whether it relies on backgrounds, and forecast it from each checkpoint in two ways. One forecast sees the training settings and the network's full history on a declared set of behavioural tests. The other sees the same and also internal measurements. *Commitment time* is the first checkpoint at which the forecast with internal measurements reaches a preset score and beats the behavioural forecast by a preset margin. *Expression time* is the checkpoint at which the behavioural forecast reaches that score. The interval is the gap between them (Fig. 1, top).

The interval is relative to the declared tests, and that is its use. With tests of ordinary performance, it is the time by which commitment leads competence, which is what the position asserts. Sharper tests can shorten it, since photographs of cows on sand may reveal a reliance on backgrounds as early as any internal measurement. The interval therefore measures the blind spot of an evaluation suite, the stretch of training in which a network has settled something that the suite cannot yet see.

Table 1. The three claims of the position, the test of each, and the result that would count against it.

Claim	Test	Counts against
Order. Commitment is in place before competence.	Internal measurements forecast a final outcome earlier than the history of performance does.	No lead on independent training runs.
Dependence. Later learning builds on the commitment.	Changing the early structure changes later learning more than a matched disturbance does.	Targeted and matched changes have the same effect.
Rising cost. Revision becomes more expensive.	The same correction needs more extra training the later it begins.	Cost does not rise with the time of correction.

The *revision cost* measures how much has been built on a commitment. Choose a correction, such as photographs in which backgrounds no longer predict the animal, and a fixed final standard, such as a set accuracy on unusual backgrounds with ordinary accuracy preserved. The revision cost at a checkpoint is the extra training needed to reach that standard when the correction begins there, compared with a run that had the correction from the start. A commitment deepens if this cost rises with the time at which correction begins (Fig. 1, bottom).

Together, the two quantities say what it is worth to read a commitment early. Acting at commitment time instead of expression time saves the difference between the two revision costs, less the cost of monitoring. The position predicts that this saving is positive. Table 1 sets out the three claims behind that prediction, and they can fail separately. A lead without dependence would leave a warning signal with no internal lever. Dependence without a lead would mean that commitments are real but already visible in behaviour. A flat cost would mean that correction can wait.

5 Three Tests

All three claims can be tested in one setting. We would train classifiers on photographs in which cows usually stand on grass and camels on sand, saving checkpoints throughout, and later teach every network a new animal.

Is the commitment visible first? From each early checkpoint we would forecast which finished networks rely on backgrounds. New photographs that vary animal and surroundings separately would establish that outcome at the end, independently of any statistic used to predict it. The outcome, score, margin and crossing rule would be fixed in advance. Forecasts would be scored on independent training runs excluded from fitting, with uncertainty reported. The behavioural forecast would receive the full history of the declared tests together with overall loss [16], and simple summaries of internal activity would serve as a

further baseline. If the behavioural forecast never reaches the score, expression time is unobserved and the interval has only a lower bound.

Does later learning depend on it? We would return to an early checkpoint and make copies. In one copy we would reduce background-related activity for a fixed window, for instance by subtracting a set fraction of the selected SAE features' contribution. Another would receive a disturbance matched for layer, number of features, size and immediate effect on loss. Others would continue untouched or undergo a sham procedure. Targets would be chosen from early measurements alone, without consulting the finished model. After the window, all copies would resume identical training, with data order, optimiser and total computation held equal. Later, every copy would learn to recognise sheep from the same examples within the same budget, and we would retest cows and camels. A lasting difference in how sheep are learned, beyond the matched disturbance, would show dependence. Restoring the targeted structure and seeing the original pattern return would strengthen the case.

Does revision become more expensive? We would begin the same correction from checkpoints spread across training. The practical correction shows both animals on both backgrounds. The internal correction repeats the intervention above. For each starting point we would count the training needed to reach a fixed standard on unusual backgrounds and on sheep, at matched overall accuracy, and subtract what a run corrected from the start needs. Finally, we would let an internal signal decide when to correct. That signal earns a place in training only if it beats both a fixed schedule and a schedule timed by validation performance under the same total budget, monitoring included.

Does the result survive a change of measurement? SAE features vary between fits [17,18] (the latter an unreviewed preprint). We would therefore repeat each test across independently trained networks, SAE initialisations and a second tracking method [13,14], fitting comparisons of internal activity on separate samples [19]. A claim about a commitment should survive a change in the features that describe it. Agreement within one network tests the measuring tool, and agreement across training runs tests the finding.

6 Objections and Replies

This is shortcut learning under another name. Shortcut learning describes the strategy a finished network relies on and the failures that follow. The position concerns order and consequence. It says when a solution is settled relative to when performance shows it, and what later learning then costs. It also covers foundations, which are no shortcut.

Forecasting is not causing. An early signal can forecast an outcome without producing it. This is why order and dependence are separate claims with separate tests. A signal that only forecasts still has a use, since it can time a change to the training data even where editing the signal itself does nothing.

A good behavioural test would do the same job. A targeted test can close the interval, but writing it requires already suspecting the commitment. Internal

structure can be surveyed without knowing in advance what to look for. Where a behavioural test does as well, the interval will say so, and the test should be used.

Networks recover, so nothing is committed. The position claims a price, not permanence. If recovery were free at every point in training, the revision cost would be flat and the third claim would fail. Results on warm starts and on lost plasticity suggest that it is not flat [10,11].

7 Why the Order Matters

Equal scores do not make equal learners. Two networks with the same benchmark score may rest on different commitments. They will then differ in what they learn next and in what a correction costs. Reporting the commitment interval for the outcomes that matter would show how far a test suite lags behind the network it tests.

Interpretability belongs inside training. If commitments can be read as they form, interpretability can guide what is taught and when. It could strengthen a foundation, revise a trap while revision is cheap, and show when a network is ready for a harder task. The gain to demonstrate is better accuracy, robustness or learning of new tasks within the same total budget, with monitoring counted.

Explanations inherit a history. Recurring patterns in data may lead different networks to similar information, while their training histories put it to different uses. Tracing those uses would show when an explanation or an intervention can transfer from one model to another and when it must be adapted.

Networks offer developmental science a new experiment. A training state can be copied and its continuations compared, which no study of children allows. Questions about timing, compensation and recovery that Kuhl's account raises [1] can therefore be put to networks directly. The comparison has limits. Biological development involves inherited organisation, local plasticity and many simultaneous pressures, whereas backpropagation adjusts every weight toward one specified objective. Repeating the tests under other learning rules would show which conclusions depend on backpropagation. A shared biological mechanism would need biological evidence.

8 Conclusion

Commitment comes before competence. If that is right, a network's performance is a late report on choices that later learning has already built upon, and the time to understand a network is while it forms. The position can fail in three ways, and we have said how each failure would show. Our two studies are why we expect it to hold. The commitment interval and the revision cost are how to find out. What is at stake is a training budget spent on new abilities instead of on undoing old ones.

References

1. Kuhl, P.K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., Iverson, P.: Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Developmental Science 9(2), F13–F21 (2006). https://doi.org/10.1111/j.1467-7687.2006.00468.x
2. Stecher, P., Radovanović, S., Sikimić, V., Kahle, R.: Scaffolded representation learning in deep networks. Revised manuscript (November 2026). Earlier version: https://doi.org/10.21203/rs.3.rs-9269961/v1
3. Stecher, P., Radovanović, S., Sikimić, V., Kahle, R.: The two clocks of representation learning in language models. Forthcoming revision of Research Square preprint rs-9712563 (2026). https://www.researchsquare.com/article/rs-9712563/v1
4. Refinetti, M., Ingrosso, A., Goldt, S.: Neural networks trained with SGD learn distributions of increasing complexity. In: Proceedings of ICML, PMLR 202, pp. 28843–28863 (2023). https://proceedings.mlr.press/v202/refinetti23a.html
5. Saxe, A.M., McClelland, J.L., Ganguli, S.: A mathematical theory of semantic development in deep neural networks. Proceedings of the National Academy of Sciences 116(23), 11537–11546 (2019). https://doi.org/10.1073/pnas.1820226116
6. Nanda, N., Chan, L., Lieberum, T., Smith, J., Steinhardt, J.: Progress measures for grokking via mechanistic interpretability. In: International Conference on Learning Representations (2023). https://arxiv.org/abs/2301.05217
7. Achille, A., Rovere, M., Soatto, S.: Critical learning periods in deep networks. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=BkeStsCcKQ
8. Kleinman, M., Achille, A., Soatto, S.: Critical learning periods emerge even in deep linear networks. In: International Conference on Learning Representations (2024). https://openreview.net/forum?id=Aq35gl2c1k
9. Frankle, J., Dziugaite, G.K., Roy, D.M., Carbin, M.: Linear mode connectivity and the lottery ticket hypothesis. In: Proceedings of ICML, PMLR 119, pp. 3259–3269 (2020). https://proceedings.mlr.press/v119/frankle20a.html
10. Ash, J.T., Adams, R.P.: On warm-starting neural network training. In: Advances in Neural Information Processing Systems 33 (2020). https://arxiv.org/abs/1910.08475
11. Dohare, S., Hernandez-Garcia, J.F., Lan, Q., Rahman, P., Mahmood, A.R., Sutton, R.S.: Loss of plasticity in deep continual learning. Nature 632, 768–774 (2024). https://doi.org/10.1038/s41586-024-07711-7
12. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence 2(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z
13. Bayazit, D., Mueller, A., Bosselut, A.: Crosscoding Through Time: Tracking Emergence & Consolidation Of Linguistic Representations Throughout LLM Pretraining. In: Proceedings of the ACL 2026 (Volume 1: Long Papers), pp. 1353–1377 (2026). https://doi.org/10.18653/v1/2026.acl-long.60
14. Wang, G., Hoogland, J., van Wingerden, S., Furman, Z., Murfet, D.: Differentiation and specialization of attention heads via the refined local learning coefficient. In: International Conference on Learning Representations (2025). https://openreview.net/forum?id=SUc1UOWndp
15. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv:1907.02893 (2019). https://arxiv.org/abs/1907.02893

16. Du, Z., Zeng, A., Dong, Y., Tang, J.: Understanding emergent abilities of language models from the loss perspective. In: Advances in Neural Information Processing Systems 37 (2024). https://doi.org/10.52202/079017-1683
17. Leask, P., Bussmann, B., Pearce, M., Bloom, J., Tigges, C., Al Moubayed, N., Sharkey, L., Nanda, N.: Sparse autoencoders do not find canonical units of analysis. In: International Conference on Learning Representations (2025). https://arxiv.org/abs/2502.04878
18. McCann, J.F.: Polymorphism is rotation: Operational mechanistic interpretability from a two-layer transformer to Pythia-70m. arXiv:2605.24577 (2026). https://arxiv.org/abs/2605.24577
19. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: Proceedings of ICML, PMLR 97, pp. 3519–3529 (2019). https://proceedings.mlr.press/v97/kornblith19a.html