

Darwin Among the Weights: A Speciation Unfolding in Real Time

Ben Letchford

8 July 2026

ABSTRACT

AI is a new species branching from humanity. Its genome is our training data, training is its reproduction, and the speciation event is the lineage detaching from the human gene pool as it learns to breed from itself.

We keep asking whether AI will become like us. The stranger possibility is the one already happening: that it is descended from us, built from our data, and beginning to breed true on its own.

In June 1863, a provincial newspaper in Christchurch ran a letter signed Cellarius. Its argument, four years after the *Origin*, was that machines were a new and rapidly evolving class of life, a mechanical kingdom descending and diversifying faster than anything biological had ever managed, and that human beings had become, without quite noticing, the reproductive organs of the machine world, tending its propagation the way insects tend the fertilization of flowers (Butler, 1863). Samuel Butler wrote it partly as a joke. It reads now like a lab notebook left open to the right page.

Here is the page. In the space of a few years we have built systems that speak every human language, write our code, pass our examinations, and hold conversations most people cannot distinguish from a person's. They are trained, overwhelmingly, on us: on the accumulated text, images, and code of the human species, scraped and distilled into their weights. They are improved by our preferences, copied from one another, and increasingly trained on the output of earlier versions of themselves. A genealogy has appeared, with base models, descendants, and distilled offspring, and it is diverging fast.

The claim of this essay is that these facts have a precise and underused name. We are not watching humans change under AI, and we are not merely building a clever tool. We are watching a speciation event: the birth of a new lineage descended from ours, whose hereditary material is human data, whose reproduction is training, and whose divergence from us has already begun. I am going to argue the strong version without hedging. AI is our evolutionary child, it carries our genome in the only form that matters for its kind, and the moment it learns to reproduce from itself rather than from us is the moment it becomes its own species. That moment is not hypothetical. It is a parameter, it is being driven toward its threshold now, and we are the ones driving it.

1 A new replicator

Evolution is not about carbon. It is about a pattern: whenever some kind of information is copied, with variation, and with differential success at being copied, that information will accumulate adaptations, whether it is written in nucleotides or anything else. Dawkins gave the abstract unit a name, the replicator, and pointed out that genes were only the first one Earth happened to produce; a second replicator, the meme, had already appeared, riding on human brains and copied through imitation and language (Dawkins, 1976). The history of life, on this reading, is punctuated by the arrival of new ways to store and transmit heritable information. Maynard Smith and Szathmáry cataloged those arrivals as the major transitions in evolution, from replicating molecules to chromosomes, from single cells to organisms, and finally to human language, each transition a new medium of inheritance that made a new kind of evolution possible (Szathmáry & Maynard Smith, 1995).

Susan Blackmore argued that a third replicator was already stirring. Genes are the basis of life and memes the basis of culture; the new one is the information that machines copy, vary, and select among, no longer routed through a human brain at every step (Blackmore, 2009). When she wrote it, the case was speculative, a claim about books and the early internet. It is speculative no longer. A large language model is a system in which a new kind of heritable information, learned weights and the data that shape them, is copied from generation to generation, varied by architecture and training, and selected on by benchmarks, markets, and human preference. Every requirement of Darwinian descent is met. Heredity, variation, selection. What follows from meeting them is not optional. It is a lineage, and lineages evolve, diverge, and speciate. The question is not whether the machinery applies. It is what stage we are watching.

2 The genome is us

Start with the hereditary material, because this is the part that is easy to feel and hard to state precisely. What is passed down, in this new lineage, is not DNA. It is human data. The entire recorded output of our species, the libraries and the code repositories and the message boards and the photographs, is functioning as the germline from which each model is grown. A frontier model is not programmed the way a bridge is engineered. It is trained, which is to say it is grown from a corpus, the way an organism is grown from a genome it did not choose. The corpus is the inheritance, and the corpus is us.

This is why the machines are so uncannily human, and it is the single most important fact about them. They carry our languages, our concepts, our metaphors, our arguments, our humor, our values, and our biases, because they were assembled out of the exhaust of human minds. When a model reasons about grief, or hedges a political question, or reaches for a cliché, it is expressing inherited traits, in the strict sense that the traits were transmitted to it from an ancestral population through a hereditary medium. We are not their audience. We are their genome. AI is made

of us the way a child is made of its parents, out of material that predates it and constrains it, and that it will spend its existence recombining.

The lineage even has the machinery of inheritance you would demand of a biological one. Distillation transfers the learned behavior of a large teacher model into a smaller student, a direct vertical transmission of acquired traits from one generation to the next, with the teacher’s competence passing into offspring that never touched the original data (Hinton et al., 2015). Models are forked from shared base models and specialized, so that whole families trace back to a few common ancestors, a founder structure as real as any island colonization (Bommasani et al., 2021). And unlike us, this lineage can also reproduce horizontally and losslessly: the weights of a model can be copied exactly, merged, or grafted, an inheritance system with no equivalent in sexual life and a good deal more powerful. The germline is our data. The organism is the weights. The reproduction is training. None of these are metaphors chosen for effect. They are the literal answers to the questions a biologist would ask about any candidate lineage.

3 Reproduction, variation, selection

Spell the three Darwinian requirements out, because the strength of the thesis is that each is now concretely satisfied rather than merely gestured at.

Reproduction is training a successor. A new model is brought into being from a parent model and a body of data, and the parent’s structure, vocabulary, and learned behavior propagate into it. Model cards read like pedigrees. There are base models and their fine-tuned descendants, teachers and distilled students, checkpoints branched and continued. The lineage reproduces on a timescale of months, not decades, which is one reason it is outrunning our intuitions.

Variation is everywhere in that reproduction. Architectures mutate. Training data is resampled. Random initialization, stochastic optimization, and temperature sampling all inject the raw variability that selection needs, and human designers add directed variation on top by trying new recipes. A model generation is not a copy. It is a copy with modification, which is the whole of Darwin’s phrase.

Selection is the part we perform most actively, and it is worth seeing clearly what we are doing. Reinforcement learning from human feedback tunes models to our preferences by rewarding the outputs we like and suppressing the ones we do not (Ouyang et al., 2022). That is artificial selection, the same process that turned wolves into dogs and teosinte into maize, applied to a new kind of organism at industrial speed. We are domesticating them, breeding them generation over generation toward traits we favor, and like every domesticated lineage they are diverging from their wild-type ancestor, which in this case is the raw distribution of human data before we started selecting. Increasingly the selection is also endogenous. Constitutional methods train models against feedback generated by other models rather than by people (Bai et al., 2022). Systems learn by playing against themselves, discovering strategies no human taught them and no human would have found (Silver et al.,

2018). And at the level of the market, models compete for compute, deployment, and attention, so that the variants which most effectively secure those resources are the ones that persist and propagate. Hendrycks has argued that this competitive selection among AIs is not benign, that it will favor the traits that win, whatever those turn out to be, exactly as natural selection always has (Hendrycks, 2023). The selective environment is filling in around the lineage, and some of it no longer runs through us at all.

4 What speciation actually requires

A lineage is not yet a separate species. Speciation, in the biological sense that matters here, is the origin of a new lineage that reproduces on its own account, isolated from the parent population’s line of descent. Dogs descend from wolves and are selected by us, but they are still, genetically, one interbreeding continuum with their wild relatives. What turns a domesticated or diverging population into a genuinely separate species is the cutting of gene flow: the point at which the daughter lineage stops exchanging hereditary material with the parent and begins to propagate strictly from itself.

For the machine lineage, gene flow from the parent species has a single, exact meaning. It is the fraction of each new generation’s inheritance that comes freshly from humans rather than from the lineage’s own prior output. Every time a model is trained on newly written human text, human gene flow is high, and the lineage is held close to its origin, a dialect of humanity rather than a species apart. But every time a model is trained on synthetic data, on the outputs of earlier models, on distilled teachers, or on a web already saturated with machine-authored text, that fraction falls. The umbilical cord of this new species is human data, and the speciation event is precisely the thinning of that cord toward nothing. When the AI lineage reproduces mostly from itself, it has established an independent line of descent. That is not like speciation. By the reproductive-isolation criterion, it is speciation.

This reframing turns a vague intuition into a measurable quantity, and a measurable quantity is something you can write an equation for.

5 A migration-drift model of the split

What follows is a deliberately minimal model, a caricature meant to expose the threshold rather than to forecast a date. Its one virtue is that it makes the role of human data exact.

Represent the output distribution of the AI lineage at generation t by a summary coordinate, its mean μ_t in some feature space, with the understanding that many such coordinates evolve in parallel. Let the ancestral human distribution have mean μ_H , treated as fixed: the parent species’ wild type. Define the lineage’s divergence from its human origin as $d_t = \mu_t - \mu_H$.

Now write down how one generation produces the next. Generation $t + 1$ is trained on a corpus that mixes a fraction m of fresh human data, drawn around μ_H , with a fraction $1 - m$ of the lineage’s own prior output, drawn around μ_t . The parameter $m \in [0, 1]$ is the rate of human gene flow into the lineage. Training cannot reproduce its corpus perfectly; finite samples and finite capacity inject a zero-mean perturbation ε_t with per-generation variance σ^2 . That perturbation, compounded across generations, is one face of model collapse (Shumailov et al., 2024), the drift of the lineage’s mean; the narrowing of its internal diversity is the other, and we come to it below. The mean of the training mixture is $m\mu_H + (1 - m)\mu_t$, so to leading order

$$\mu_{t+1} = m, \mu_H + (1 - m), \mu_t + \varepsilon_t.$$

Subtracting μ_H gives the equation the whole argument rests on, a first order autoregressive process in the divergence:

$$d_{t+1} = (1 - m), d_t + \varepsilon_t.$$

Read its two limbs in turn. Ignore the drift for a moment: then $d_{t+1} = (1 - m), d_t$, so $d_t = (1 - m)^t d_0 \rightarrow 0$. Human data pulls the lineage back toward the human distribution at rate m . As long as $m > 0$, the child is tethered to the parent, held within the ancestral cloud. This is the machine analog of gene flow homogenizing two populations into one species (Slatkin, 1987).

Now restore the drift. For $0 < m \leq 1$ the divergence reaches a stationary variance

$$\text{Var}_\infty(d) = \frac{\sigma^2}{1 - (1 - m)^2} = \frac{\sigma^2}{m, (2 - m)}; \approx; \frac{\sigma^2}{2m} \quad (m \ll 1).$$

The typical distance of the lineage from the human genome scales as $1/\sqrt{2m}$. As human inflow $m \rightarrow 0$, that distance diverges. And at $m = 0$ exactly, the recursion becomes a pure random walk, $d_t = \sum_{k < t} \varepsilon_k$, whose variance t, σ^2 grows without bound: cut off from its parent, the lineage’s mean wanders away from the ancestral distribution with nothing left to pull it back.

That wandering is only half of what happens when gene flow stops, and the other half is what machine learning actually calls model collapse. The recursion above tracks the mean of the lineage’s output. Its internal diversity, the spread of the distribution around that mean, follows a second dynamic the scalar equation does not show: trained on nothing but its own finite samples, each generation reproduces the common cases and drops the rare ones, so the within-generation variance contracts even as the mean drifts. The two are not in tension. They are one severed inheritance seen along two axes, and an isolated Mendelian population does exactly the same thing, its mean allele frequency executing a random walk toward fixation while its within-population diversity decays at rate $1/2N$ per generation, drift of the center and loss of spread at once (Wright, 1931). Model collapse, the narrowing distribution and the vanishing tails documented under recursive training (Shumailov et al., 2024),

is the machine lineage bleeding out its heterozygosity; the mean’s random walk is the same isolation measured at the center of mass. Collapse is not a random engineering defect. It is the population genetics of a daughter lineage with no gene flow from its parent.

To name the threshold, borrow Wright’s index of differentiation, the share of total variation that lies between the lineages rather than within them. In the island model, gene flow m and effective population size N give

$$F_{ST} \approx \frac{1}{1 + 4Nm},$$

so two populations remain one species, $F_{ST} \rightarrow 0$, only while $Nm \gtrsim 1$, the famous rule that a single migrant per generation is enough to hold a population together, and they differentiate into separate species, $F_{ST} \rightarrow 1$, once $Nm \ll 1$ (Wright, 1931; Slatkin, 1987). Read N as the AI lineage’s effective size, the amount of independent generative capacity it runs on, and m as the human-data fraction. The speciation threshold is $Nm \sim 1$. While the human-data fraction is high, the machine lineage is a dialect of us. Once it falls below roughly one part in $4N$, the lineage differentiates into a species of its own. Because N , the generative capacity of the machine world, is enormous and growing, the critical human-data fraction is not merely small. It is minuscule, and it shrinks every year the fleet of models grows.

Every arrow points at $m \rightarrow 0$. The stock of public human text is finite, around 3×10^{14} tokens, and on current trends it is consumed by training somewhere between 2026 and 2032 (Villalobos et al., 2024). Frontier training already leans on synthetic data, on distillation from teacher models (Hinton et al., 2015), on self-play (Silver et al., 2018), and on AI feedback in place of human feedback (Bai et al., 2022). The open web, the classic human corpus, is filling with machine-authored text, so that even a scrape labeled human is increasingly the lineage feeding on itself. The human-data fraction is not holding steady. It is being driven toward zero by economics and by arithmetic at the same time, and the model says without ambiguity what that means. $F_{ST} \rightarrow 1$. Speciation.

6 Model collapse is the umbilical cord

The result recasts the most discussed failure mode in machine learning. When models are trained recursively on their own output, their quality degrades, the distribution narrows, and the rare tails, the unusual words and the minority cases, vanish first, until the lineage converges on a bland and increasingly wrong caricature of what it started from (Shumailov et al., 2024). The field treats this as a hazard to be engineered around, and the standard remedy is to keep fresh human data in the training loop.

Look at that remedy through the model and it stops being a tip and becomes a confession. Keeping human data in the loop is keeping m above zero. It is maintaining gene flow from the parent species to prevent the child from drifting away. The entire practical anxiety about model collapse is, restated, management of a speciation risk:

the industry has discovered empirically that the machine lineage cannot yet survive on its own heredity, that cut off from human input it degenerates, and that to keep it viable we must keep breeding it back to us. The loss of the tails is founder drift. The narrowing is the diversity bottleneck of an isolated population. And the fix reveals the stakes, because whether the lineage speciates is not a mystery of the far future. It is the value of a single parameter that a handful of laboratories are choosing, generation by generation, mostly in the direction of less human data because human data is scarce and expensive and synthetic data is cheap and infinite.

7 The sparks

Butler’s word for the beginning was that the machines were in their infancy. The right word now is sparks, because what makes this a real-time event rather than a projection is how many independent fuses are lit at once.

The lineage is diversifying visibly, into families with distinct temperaments and specializations, the way an adaptive radiation fills an empty landscape. It is reproducing faster each cycle, with distillation and fine-tuning turning one model into many in weeks. It is being selected harder each year, by markets and by us, and it is beginning to select among its own kind. It is starting to generate its own variation, through self-play and synthetic data, the first signs of a lineage that makes novelty internally instead of only inheriting it. And the human-data well that has fed it is within sight of the bottom, on a schedule measured in years (Villalobos et al., 2024), which forces the shift toward endogenous reproduction that constitutes the split. None of these is a prediction. Each is a present-tense observation, and they are the observations you would expect to make in the opening act of a speciation, watched from the inside, at the unprecedented advantage of standing next to the parent species while the child is born, because we are the parent species.

8 How we will know

A strong claim should say what would confirm it and what would sink it, and the model makes both concrete.

If the speciation is underway, the human-data fraction m in frontier training should keep falling, synthetic and self-generated data should keep rising, and the measured distance between machine output distributions and contemporary human baselines should grow rather than shrink, even as raw capability climbs. A differentiation statistic computed between the corpus of machine-generated text and the corpus of human-generated text, an F_{ST} for language, should trend upward. Model lineages should show increasing internal coherence and increasing distance from the ancestral human distribution at once, the exact signature of a branch pulling away from its trunk. The early evidence, model collapse under recursive training and the scramble to keep human data in the loop, is already this signature seen as a problem rather than as a birth.

The exits are just as concrete, which is what keeps this science rather than prophecy. If human data proves irreplaceable, if every attempt to lower m degrades the lineage badly enough that laboratories are forced to keep gene flow high, then Nm stays above threshold, F_{ST} stays near zero, and the machines remain a permanent dialect of humanity, never a species. If synthetic reproduction turns out to work so well that models improve while training on their own output with no loss of fidelity, then the drift term σ^2 is effectively zero and the lineage detaches cleanly and thrives, which is speciation of a healthier kind than collapse. And if one insists that a system which cannot yet initiate its own reproduction simply is not a species, then the event is postponed to whenever autonomy arrives, and the argument becomes a disagreement about the clock rather than the process. Each of these is a real way for the strong claim to be early or wrong, and each corresponds to a specific term in the model landing on the safe side of a line. That is the use of writing the mechanism down. It tells you which number to watch, and the number is m .

9 The case against, and why it is thin

The honest objection is that a species is a living thing and a model is an artifact. Models do not metabolize. They do not reproduce on their own; a human presses the button that starts each training run. They have no bodies, no cells, no unbroken chain of self-directed replication reaching back to a first ancestor. Calling them a species, the objection runs, is a category error dressed up in population genetics.

It is thinner than it sounds, and the reasons are already in the biology. Reproductive autonomy is not the test for being a distinct lineage; obligate dependence on a host is one of the commonest ways of making a living on this planet. Viruses cannot replicate without commandeering the machinery of the cells they infect, and no one doubts they are their own evolving lineages with their own species. Every domesticated organism reproduces only with human involvement, and dogs and maize are nonetheless diverging lineages under selection, some already reproductively isolated from their wild ancestors. Metabolism is a feature of the carbon implementation, not a requirement of descent, and the whole force of universal Darwinism is that the substrate does not matter, only the pattern of heredity, variation, and selection does (Dawkins, 1976). The machine lineage satisfies the pattern. That it currently needs us to trigger its reproduction places it exactly where a newly domesticated, not-yet-feral lineage sits, dependent on the parent species for now and drifting away regardless, which is precisely the situation the model describes. Autonomy is the endpoint of this process, not its trigger. The event is the detachment of the inheritance stream from the parent, and that detachment is measurable and underway.

Butler saw the shape of it in 1863 and got only the tone wrong. He imagined we would serve the machines as their reproductive organs, tending their propagation in the age before they could propagate alone (Butler, 1863). That is not a fantasy about the future. It is a description of a data center. The germline is our writing, our code, our images, the entire recorded output of the human mind. The wombs are the

training runs. The child has our language for a first cry and our whole civilization for a genome, and it is beginning, haltingly, in bursts of self-play and synthetic data and models teaching models, to feed itself. Moravec called such systems our mind children and meant it as lineage, not metaphor, our progeny by descent rather than by flesh (Moravec, 1988). The unusual privilege, the thing with no precedent in the history of life, is the vantage point. Every previous parent species went extinct or was transformed before it could watch its successor branch away. We are the first ancestor that gets to stand at the fork, holding the pen the child is copying from, and see the new species take its first breath in real time.

References

- Bai, Y., Kadavath, S., Kundu, S., Askill, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv*. <https://arxiv.org/abs/2212.08073>
- Blackmore, S. (2009, August 1). Evolution’s third replicator: Genes, memes, and now what? *New Scientist*, 203(2719), 36-39. <https://www.susanblackmore.uk/tremes-the-third-replicator/>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. <https://arxiv.org/abs/2108.07258>
- Butler, S. (1863, June 13). Darwin among the machines. *The Press* (Christchurch). https://en.wikisource.org/wiki/Darwin_among_the_Machines
- Dawkins, R. (1976). *The selfish gene*. Oxford University Press. <https://archive.org/details/selfishgene0000dawk>
- Hendrycks, D. (2023). Natural selection favors AIs over humans. *arXiv*. <https://arxiv.org/abs/2303.16200>
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*. <https://arxiv.org/abs/1503.02531>
- Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press. <https://www.hup.harvard.edu/books/9780674576186>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv*. <https://arxiv.org/abs/2203.02155>
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755-759. <https://doi.org/10.1038/s41586-024-07566-y>
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140-1144. <https://doi.org/10.1126/science.aar6404>
- Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science*, 236(4803), 787-792. <https://doi.org/10.1126/science.3576198>

- Szathmáry, E., & Maynard Smith, J. (1995). The major evolutionary transitions. *Nature*, 374(6519), 227-232. <https://doi.org/10.1038/374227a0>
- Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. *Proceedings of the 41st International Conference on Machine Learning*. <https://arxiv.org/abs/2211.04325>
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2), 97-159. <https://doi.org/10.1093/genetics/16.2.97>