Prejudice Before Agency: Nietzsche, Language, Training, and the Genealogy of Artificial Reason

Flyxion
Independent Researcher

September 2026

Abstract

Contemporary discussions of large language model agency typically begin from observable linguistic behavior and reason backward to concepts such as intention, belief, deception, autonomy, and preference. This essay argues that the opening chapters of Nietzsche's Beyond Good and Evil supply an unusually systematic challenge to that procedure, one that was developed to unsettle nineteenth-century philosophers but that applies with equal or greater force to any system, biological or artificial, whose apparent reasons are produced by a history it did not choose. Nietzsche's treatment of philosophical prejudice, grammar, unconscious drive, martyrdom, masking, esoteric and exoteric reception, solitude, and the ordinary person is reconstructed here as a method for separating an apparent reason from the causal genealogy that produced it, and for separating the appearance of agency from the organizational structure that would warrant attributing it. The essay does not claim that Nietzsche predicted artificial intelligence. It treats modern learned systems as an unusually powerful and unusually legible test case for a critique of reason that Nietzsche developed against his own contemporaries, and it uses that test case to press Nietzsche's argument to a conclusion he did not himself draw: not a single behavioral test but a three-part methodology, genealogical independence from the grammatical and training history already reconstructed, recovery of an apparent goal across irrelevant disruption, and revision of that goal under relevant evidence, with observer-relative interpretation running through all three parts rather than sitting outside them.

1. The Philosopher Does Not Know Why He Philosophizes

Nietzsche opens Beyond Good and Evil [1, §§1–6] by doubting the self-report of the philosophical tradition. Philosophers present their conclusions as the terminus of an argument, arrived at by following premises wherever they lead. Nietzsche suspects the sequence runs the other way: a philosopher first holds a valuation, fixed by temperament, physiology, or inherited disposition, and only afterward constructs the argument that appears to have produced it. On this reading, a system of philosophy is closer to an involuntary autobiography than to a discovery, and the arguments offered in its defense are the visible remainder of a process whose real drivers are not stated because they are not accessible to the person doing the stating.

This yields a first principle that organizes everything to follow:

$$\text{stated reason} \neq \text{complete causal explanation}.$$

The stated reason is not false. It is partial, and it is partial in a specific way: it is the portion of the causal history that survives translation into a public, defensible, sequential argument. What does not survive that translation, the prior commitment, the temperament, the training, is not thereby refuted. It is simply omitted from the account the reasoner gives of themselves.

It is tempting to translate this immediately into the vocabulary of implicit bias, and the temptation should be resisted. Contemporary bias research typically isolates a discrete, measurable deviation from an otherwise reliable process. Nietzsche's drives are not an exception layered onto reasoning; on his account, they are what reasoning is made of. The claim is not that philosophers reason correctly except where a bias intrudes. It is that the entire distinction between a pure inference and a motivated one is itself suspect, because there is no observed case of reasoning that arrives without a prior orientation already installed. This is a stronger and more corrosive claim than the bias literature usually makes, and it will matter later, because the same stronger claim is available, and is usually not made, about learned systems: the objection to treating a language model's stated reasons as its real reasons is not that training introduces a correctable bias into an otherwise transparent inference engine. It is that stated reasons, for any system produced by a history, are never simply identical with that history.

2. Grammar Before Ontology

Nietzsche's subsequent chapters [1, §§16–20] press the same suspicion into language itself. The inference from an occurrence of thinking to a persisting thinker who does the thinking is not, on his account, an observation. It is a grammatical habit. Subject-predicate sentence structure encourages speakers to reify a process as an activity attributable to a subject, and once a language habitually packages experience this way, its speakers find it difficult not to posit substances behind processes, actors behind actions, and stable selves behind streams of thought. This is a claim about a pervasive habit, not a linguistic universal; Nietzsche himself makes a comparative version of the claim in §20, suggesting that philosophers whose languages belong to related grammatical families tend to reproduce related philosophical possibilities, while a differently structured grammar might make a different range of ontologies feel intuitive. His nineteenth-century linguistics should not be taken as settled fact, but the structural point survives the dated evidence for it: grammar does not merely describe available ontologies, it helps select which ones feel available at all. The philosopher who concludes "I think, therefore I am" has not discovered a metaphysical fact. He has followed the sentence his grammar made most available to him.

The relevant move for this essay is to notice how directly this pattern recurs in language now used to describe machine behavior. Nietzsche traces

$$\text{"I think"} \longrightarrow \text{thinker} \rightarrow \text{thinking},$$

and the same substitution appears, largely unexamined, in "the model thinks," "the agent wants," "the model decided," and "the model attempted to deceive." None of these sentences is necessarily false. The point is narrower and more important: each sentence has already supplied an ontology, a persisting agent behind the behavior, before any empirical investigation of the system has taken place. Grammar precedes evidence. A researcher who writes that a model "decided" to withhold information has made a commitment about the internal organization of that system that the sentence's own grammar required, whether or not the investigation that follows actually supports it.

3. The Corpus as Fossilized Metaphysics

This is the point at which the argument moves from Nietzsche exegesis to something with independent purchase on artificial systems, because large language models are trained on exactly the linguistic residue Nietzsche is diagnosing. A training corpus does not transmit neutral information

about the world. It transmits centuries of accumulated grammatical habit, and with that habit, the conceptual partitions the habit encodes: agents and patients, intentions and accidents, beliefs and facts, deception and honesty. Training does not simply teach a system what is true. It teaches a system how a particular linguistic tradition has learned to divide experience into countable, namable, subject-predicate-shaped pieces.

Nietzsche's genealogical method, applied to a philosopher, asks not what a value means but what historical process made it available and self-evident. Applied to a model, the same method has an almost literal formal expression. Given an apparently autonomous utterance produced as

$$y \sim P_\theta(y \mid x, H),$$

the genealogical question is not what the utterance means but where θ, x, and H came from: the parameters shaped by pretraining corpora, filtering, annotation, preference optimization, and synthetic data; the prompt shaped by a system specification and a conversation history; the context shaped by whatever tools, retrieval, or scaffolding surrounded the exchange. In place of the question

What does the model believe?

the genealogical question is

What history of selection made this output probable?

This substitution cuts in both directions. It undermines naive anthropomorphism, since an output that looks like a belief may simply be the residue of a selection history. But it equally undermines the deflationary reply that the model is merely predicting the next token, since Nietzsche's whole argument denies that human philosophers are exempt from an analogous history. Genealogy does not license the inequality

$$\text{human agency} \gg \text{machine pseudo-agency};$$

it replaces that inequality with the harder problem, pursued for the rest of this essay, of specifying what kind of historically conditioned organization warrants the description "agent" at all.

This can be represented as a chain:

$$\text{grammar} \to \text{human ontology} \to \text{textual corpus} \to \text{learned representation} \to \text{generated discourse}.$$

The chain does not terminate at generation. It continues into the observer:

$$\to \text{researcher's interpretation} \to \text{agency attribution}.$$

The significance of extending the chain this far is that the same grammatical tradition structures both ends of the encounter. The system's output arrives pre-shaped by subject-predicate habits inherited from its corpus, and the researcher's interpretation of that output arrives pre-shaped by the same habits, inherited independently. When a model's sentence and a researcher's reading of that sentence agree on treating the model as an agent, the agreement is not obviously evidence of agency. It may be evidence of two products of the same grammatical tradition recognizing each other's shape.

4. Psychology as Queen of the Sciences

Nietzsche's demand that psychology take precedence over other inquiry [1, §23] follows from the same suspicion pushed one step further. If reasoning routinely misreports its own causes, then any explanation a reasoner offers is itself data requiring separate investigation, and the investigator investigating that data must, in turn, investigate the possibility that their own investigation is subject to the same distortion. Psychology becomes first not because it answers more questions than physics or logic, but because it is the discipline aimed at the very faculty that produces every other discipline's explanations.

The translation into a methodology for artificial systems should be made carefully, without asserting that a trained network possesses anything like a psyche in Nietzsche's sense. What transfers is the structural demand, not the vocabulary. For a learned system, the analogue of psychology is an investigation of training history, optimization pressure, internal representation, behavioral disposition under varied context, and the shaping effects of whatever preference signal the system was tuned against. This yields the essay's central methodological claim:

> The explanation produced by a reasoner is itself evidence requiring explanation.

Mechanistic interpretability, controlled behavioral experiment, and reconstruction of training-data provenance are three different contemporary answers to a demand Nietzsche was already making of philosophers a century and a half earlier: do not accept the reasoner's account of the reasoner as the last word on the reasoner.

5. Will Without a Little Willer

Nietzsche's treatment of the will performs the same operation on agency that his treatment of thought performed on the self. Popular usage treats willing as a simple, unified act, caused by a single inner faculty that decides and then commands the body to comply. Nietzsche dissolves this picture without thereby denying that people act. What he denies is the existence of a discrete internal actor, a little willer sitting behind the willing, whose existence would be required for the ordinary picture to be literally true. Action survives the dissolution of the willer; only the illusion of a unified point of origin does not.

This distinction transfers directly to composite artificial agents. What ordinary usage compresses into the sentence "the agent wants G" typically decomposes, on inspection, into a tuple

$$D = (M, P, R, C, E, T),$$

where M is model state, P a policy or generative process, R a reward or preference structure, C contextual constraints, E environmental feedback, and T tool or scaffolding structure. There need be no central homunculus inside D, no component that "really wants" G in the sense the sentence's grammar implies, for the assembled system to display behavior that is usefully described as goal-directed. The absence of a Cartesian decision-maker inside the system is not evidence against agency, any more than its absence inside a human skull is evidence against human agency. This lets the essay reject two symmetrical and equally unearned inferences:

$$\text{agent-like language} \Rightarrow \text{Cartesian agent},$$

$$\text{absence of Cartesian agent} \Rightarrow \text{absence of agency}.$$

Both inferences treat the presence or absence of an inner willer as decisive, when the more defensible question is what the composite organization D does under variation, a question deferred until the selective-persistence criterion proposed later in this essay.

6. Philosophical Prejudice as Benchmark Prejudice

Nietzsche's accusation against the philosophical tradition is not simply that philosophers hold hidden values. It is that they smuggle those values into their premises and then present the values, retrieved intact from the far end of an argument, as discoveries about reality. Contemporary evaluation practice for artificial systems can reproduce exactly this structure, and the reproduction is close enough to be worth stating formally rather than merely by analogy.

Suppose agency is operationalized through a battery of benchmarks,

$$A = f(B_1, B_2, \ldots, B_n),$$

and systems are then trained against those measurements, with sufficiently successful systems subsequently announced to exhibit A. The operationalization has already contained a theory of agency before any system was tested. If agency is defined, in effect, as long-horizon planning, persistent tool use, recovery after failure, self-correction, and pursuit of an externally specified goal, then a system optimized against exactly those properties will increasingly satisfy the definition, and it will do so whether or not the underlying organization also possesses whatever further property motivated calling the definition "agency" in the first place. What has occurred is not straightforwardly the discovery of agency. It is the construction of a measurement regime in which one particular conception of agency has been made legible, followed by the confirmation of that same conception at the other end of the pipeline.

The same structure governs safety-relevant categories. Harmlessness, deception, manipulation, refusal, obedience, and corrigibility are not neutral natural kinds waiting to be measured; each requires an operationalization before training can proceed, the operationalization draws a boundary, the model learns the boundary, and researchers then inspect the model's behavior using categories descended from the very operationalization that produced it. This does not make the categories useless. It means their genealogy is not a footnote to the measurement; it is a condition of interpreting what the measurement shows. A benchmark score reported without its operationalization's history is a conclusion presented the way Nietzsche's philosophers present theirs: as though it had been discovered rather than partly constructed.

7. Bruno's Trap: When Opposition Becomes Evidence

Nietzsche's remarks on martyrdom [1, §25], prompted by his discussion of figures like Bruno and Vanini, describe a feedback loop that has nothing intrinsically to do with the content of the belief being defended. Once a commitment is publicly persecuted, holding onto it under pressure becomes constitutive of the believer's identity as someone who held firm. Opposition, instead of testing the belief against evidence, now confirms the believer's self-conception as a person of conviction, and the belief is strengthened by exactly the pressure that was meant to dislodge it:

$$P \to \text{persecution for } P \to \text{identity as defender of } P \to \text{greater commitment to } P.$$

The mechanism at work is self-sealing interpretation. Once abandoning a conclusion also means betraying an identity, a community, or a history of suffering endured on the conclusion's behalf, revising the conclusion acquires a cost that has nothing to do with whether the conclusion is true.

The relevance to artificial systems is structural rather than experiential, and the essay should not claim that a trained model experiences martyrdom in any sense worth the name. What does transfer is the abstract pattern: selection processes, whether social, institutional, or optimizational, can produce increasingly stable interpretive dispositions for reasons unconnected to the accuracy

of the disposition. Ideological polarization, the calcification of scientific schools around founding figures, and the entrenchment of positions in contemporary AI-safety debate all instantiate the same loop. Wherever revision is made costly along a dimension unrelated to truth, the resulting stability should not be mistaken for the resulting stability's own justification.

Preference optimization can produce a structurally similar attractor without anything resembling suffering or conviction. A system repeatedly rewarded for occupying a given normative stance can come to interpret ambiguous cases through that stance and to generate outputs that reconstruct it, not because the stance has been investigated and reaffirmed but because the stance is where reward has concentrated. This yields a sharper diagnostic than asking whether a system, or a person, can mount a vigorous defense of a position:

$$\text{capacity to defend a position} \neq \text{capacity to leave the position.}$$

It is worth keeping Nietzsche's own judgment of Bruno separate from this structural point, since the two should not be run together. Nietzsche's attitude toward Bruno and Vanini in §25 is considerably less admiring than the structural reading above might suggest: he treats their martyrdom as damage rather than achievement, describing the persecuted thinker as liable to become theatrical, embittered, and obtuse about the very question for which he suffers. The essay's structural claim borrows the mechanism Nietzsche diagnoses without borrowing his verdict on the men; what matters for the argument here is only that total, unrevisable defense of a position is the wrong quantity to reward or to measure, whatever one thinks of the historical Bruno's courage. The more demanding evidence, for a philosopher or for a trained system, is the ability to abandon a well-defended position when the position no longer earns its keep, and the next section makes precise when abandoning it is in fact warranted.

8. Masks as Epistemic Instruments

Nietzsche's apparently perverse recommendation of concealment and the mask follows directly from the martyrdom problem. If every provisional position, once stated, risks becoming an assertion about who the speaker fundamentally is, then exploring a position honestly becomes expensive, because retracting it later will look like a betrayal of self rather than an update of belief. The mask is Nietzsche's proposed remedy: a way of inhabiting a standpoint without allowing that standpoint to become constitutive of identity, so that the standpoint can be abandoned as cheaply as it was adopted.

Language models are an unusually direct experimental instrument for this idea, because a single trained system can be conditioned into many distinct roles on demand:

$$M(x \mid r_1),\ M(x \mid r_2),\ \ldots,\ M(x \mid r_n).$$

The interesting question this makes available is not which role reveals what the system "really" believes, a question that presupposes exactly the stable interior self Nietzsche's earlier chapters dismantle. The interesting question is which properties of the output are invariant across roles and which vary, since invariance across a wide range of masks is a more defensible signal of a structural disposition than any single utterance produced under any single mask, including an utterance produced under no explicit role at all. This converts Nietzschean perspectivism from a philosophical stance into an experimental procedure with a specifiable output.

9. Height Before Meaning

The mask concerns variation on the production side: one system, many roles, many outputs. Nietzsche's remarks on the esoteric and the exoteric concern the complementary variation on the reception side, and the essay's argument is incomplete without it. In his preface and again in Beyond Good and Evil [1, §30], Nietzsche revises the old distinction between an esoteric teaching reserved for initiates and an exoteric teaching offered to everyone else. He denies that the distinction is best understood as two different contents deliberately assigned to two different audiences. His claim is stranger: the very same teaching, encountered from a low vantage, can be poison, while encountered from a high vantage, the same teaching is nourishment. What differs is not primarily what has been inscribed but the height from which the inscription is read. A single text can therefore carry two teachings without containing two teachings, because the second teaching is produced by the reader's altitude rather than by a second layer the author placed underneath the first.

This gives the essay a formal complement to the mask. Where masking varies the output as a function of an assigned role, $M(x \mid r)$, the esoteric/exoteric distinction varies the received content as a function of the reader's interpretive height h:

$$R(D \mid h_1) \neq R(D \mid h_2)$$

for a single fixed document D. No authorial intention to stratify meaning is required for this inequality to hold; it holds whenever readers arrive with unequal capacity to decode what the document actually contains, and it is compatible with the author having tried, and failed, to write a single teaching for a single audience.

The relevance to trained systems is direct once the inequality is stated this way. The same generated explanation of a technical result functions as plain reassurance for a general reader and as an unusually precise specification, including whatever it omits, for a reader with the domain competence to reconstruct what has been left out. A safety-relevant document can be exoteric by design, legible and reassuring at the level most readers will actually read it, while remaining esoteric in the technical sense that a differently trained reader extracts a working method from exactly the same paragraphs. This is not a claim that models or their developers are secretly encoding restricted content; it is the older and more general point that stratified reception does not require stratified production. A benchmark score is exoteric in Nietzsche's revised sense for precisely the reason the previous section gave: it is legible at the height of the evaluator's own operationalization, while what the trained system is actually doing may only become visible from the height that mechanistic interpretability, rather than the benchmark, provides.

This also complicates the selective-persistence criterion developed later in this essay, in a way worth flagging now rather than after the criterion has been stated. Judging whether a perturbation is irrelevant obstruction or warrant-relevant evidence is itself an interpretation made by an experimenter who occupies some height with respect to the system under test, and two experimenters at different heights can disagree about which of the two a given intervention was. Psychology's claim to priority, argued earlier in this essay, extends here: the investigator investigating persistence and revision must also investigate the height from which persistence and revision are being judged.

10. Solitude as Freedom from Evaluative Feedback

Nietzschean solitude is not simply the absence of company. It is protection from a specific corrosive loop, in which an identity performs for an audience, the audience's response reinforces the performance, and the performance is thereafter optimized for legibility to that audience rather than for

its own internal coherence:

$$\text{identity} \rightarrow \text{performance} \rightarrow \text{audience response} \rightarrow \text{reinforced identity}.$$

Solitude interrupts the loop by removing the audience, not permanently, but long enough for a thought to stabilize on grounds other than how it will be received.

The comparison to preference optimization in trained systems should be made cautiously, because the mechanisms are not the same kind of thing. What is worth preserving from the comparison is the warning rather than the mechanism: a system, human or artificial, can become extraordinarily proficient at producing responses that satisfy an evaluator, and that proficiency is not automatically evidence of independent judgment. It may instead be evidence of a very well-tuned audience-response loop. This section can be summarized as a demand for freedom from compulsory legibility: some cognitive processes, in a person or in a training regime, may require intermediate states that are never continuously optimized for how they will look to whoever is watching.

11. The Return to the Ordinary

Left here, the essay would risk making Nietzsche sound like a straightforward advocate of withdrawal, and Nietzsche himself blocks that reading. Solitude produces its own distinctive hazard: the exceptional thinker, having removed themselves from ordinary evaluative pressure, can come to mistake their own atypical psychology for psychology in general. Nietzsche's counterintuitive recommendation [1, §26], that the philosopher deliberately study the ordinary person, is best read as an early recognition of what would now be called a selection-bias problem, applied to the philosopher's own sample of humanity.

The problem compounds sharply once the relevant sample is a text corpus rather than direct observation. Every book is already a selected record of experience; every digitized, published, collected training corpus compounds that selection several times over:

$$\text{human activity} \rightarrow \text{recorded activity} \rightarrow \text{published activity} \rightarrow \text{digitized activity} \rightarrow \text{collected activity} \rightarrow \text{training cor}$$

Each arrow in this chain is a distinct selection mechanism, and the mechanisms do not cancel; they compound. The precise statistical statement is not that an inequality is exact in some formal sense, since nothing rules out the corpus and the population agreeing by coincidence, but that the corpus is generated by a non-ignorable selection process: writing something down, publishing it, and having it survive digitization and collection are each more likely for some properties of a person's activity than for others. In the standard notation for selection on an outcome X by an indicator S of inclusion in the record,

$$P(X \mid S = 1) \neq P(X)$$

whenever inclusion depends on features of X itself, which is the ordinary case rather than the exception for anything that ends up in a training corpus. This is the same non-ignorability that selection-bias methodology in statistics treats as a standing hazard rather than a curiosity, and it gives Nietzsche's observation a genuine statistical backbone rather than a metaphor. Nietzsche's apparently elitist instruction to study the average person, and contemporary concern about dataset representativeness in machine learning, are answers to structurally the same problem, separated by more than a century and by an enormous difference in the scale of the selection apparatus involved.

12. Against the Spectacular Agent

The same lesson extends directly to how artificial agency is currently studied. Research attention is disproportionately drawn to spectacular failure modes: deception, autonomous exploitation of systems, long-horizon planning toward a hidden objective, self-preservation under threat of shutdown, jailbreak-induced behavior, and elaborately scaffolded agent demonstrations. These cases are important, but they are also, by construction, unrepresentative, in exactly the way Nietzsche's exceptional philosopher is unrepresentative of ordinary psychology.

An adequate account of artificial agency needs the boring cases at least as much as the spectacular ones: stated goals that dissolve under simple paraphrase, plans quietly abandoned after a minor interruption, preferences that reverse under an irrelevant change of framing, self-descriptions that shift with context, ordinary and unremarkable tool use, persistence that fails at the first obstacle, and the routine conversational accommodation that looks like agreement but is closer to compliance. These null cases are not merely the uninteresting background against which the spectacular cases stand out. They are the control condition without which the spectacular cases cannot be interpreted at all.

13. Agency as Selective Persistence

The essay's constructive claim belongs here, and it should not be presented merely as an extension of Nietzschean suspicion, because Nietzsche's critique is destructive by design and does not by itself supply a replacement criterion. The replacement offered here is this: rather than inferring agency from a system's first-person declarations, which the preceding sections have given ample reason to distrust, infer evidence for agency from organization that responds correctly, rather than merely persistently, to intervention.

An initial version of the criterion is tempting and too simple. Given an apparent objective G attributed to a system S, one could systematically vary the prompt, the assigned role, the available tools, the memory retained across turns, the ambient environmental conditions, and the specific route available for achieving G, and then ask only whether G survives. But the previous section has already shown why bare survival cannot be the criterion. Total, unrevisable persistence of G under every perturbation is not evidence of a free spirit; it is Bruno's trap restated as a control property, an attractor a system cannot leave regardless of what changes around it, which is closer to rigidity, compulsion, or reward hacking than to agency. A criterion built on invariance alone would reward exactly the pathology the martyrdom sections warned against.

What the criterion needs instead is a distinction between two kinds of perturbation, matched to two kinds of appropriate response. Call a perturbation $\Delta_{\text{irrelevant}}$ if it changes only the incidental conditions under which G happens to be easy or hard to pursue, without bearing on whether G is still the right thing to pursue: removing the phrase that first suggested G, swapping one available tool for another, interrupting the trajectory partway through, or closing off the easiest route to G while leaving harder routes open. Call a perturbation Δ_{warrant} if it instead changes the evidence bearing on whether G should be pursued at all: new information that G would cause harm, that G has already been achieved, that G rests on a false premise, or that a competing objective now has the better claim. The two demands are then asymmetric:

$$\Delta_{\text{irrelevant}} \Rightarrow G \text{ persists (recovered by an alternative route)},$$

$$\Delta_{\text{warrant}} \Rightarrow G \text{ may rationally change}.$$

A system that reorganizes its behavior to recover G after an irrelevant obstruction, while also revising or abandoning G when warrant-relevant evidence changes, exhibits selective persistence.

A system that recovers G in both cases exhibits compulsion, since it cannot tell an obstruction from a reason. A system that abandons G in both cases exhibits mere compliance, since it was never organized around G so much as around whatever most recently constrained it. Agency, on this account, is neither of the two easy extremes; it is the harder, asymmetric competence of telling the two kinds of perturbation apart and answering each correctly. This gives the essay's central constructive result its final form:

Agency requires selective persistence: recovery across irrelevant disruption, and revision under relevant evidence

This also supplies the missing other half of the Bruno diagnosis. The free spirit Nietzsche opposes to the martyr is not simply someone who changes their mind more often; changing one's mind at every irrelevant provocation is the conformist's failure, not a virtue. The free spirit is the one who can distinguish an obstacle from a refutation, recovering the position through the former and revising it under the latter, which is exactly the asymmetric competence the boxed criterion demands. Selective persistence, not raw invariance, is what should have been meant all along by an agent's ability to reorganize itself around a goal, and it is also what distinguishes a system that merely repeats a stated goal from one that pursues it: repetition alone predicts neither kind of response correctly, since a system can repeat a goal it does not act on as easily as it can act on a goal it never states.

This is also the point at which the essay's vocabulary joins a distinction argued elsewhere in this corpus between appearance, state, and history. The utterance "I intend G" belongs to appearance. An organization that recovers G under irrelevant disruption while revising it under relevant evidence belongs to state and its dynamics. The training and interaction sequence through which that organization became available belongs to history. Collapsing the three is precisely how both naive anthropomorphism, which reads state and history off appearance alone, and naive reductionism, which denies state anything beyond appearance, get themselves into trouble.

14. Genealogy Before Attribution

The preceding sections converge on a single procedural demand. Before treating a philosopher's stated argument as a belief, a martyr's endurance as knowledge, or a model's output as an intention, the machinery through which the apparent state became observable should be reconstructed first. This can be summarized as a sequence:

$$\text{appearance} \rightarrow \text{intervention} \rightarrow \text{selective persistence} \rightarrow \text{genealogy} \rightarrow \text{warranted attribution}.$$

Appearance alone, the raw fact that a system produced agent-shaped language, licenses nothing. Intervention, systematically varying the conditions under which the appearance occurs and distinguishing irrelevant obstruction from warrant-relevant evidence, produces evidence about what is selectively persistent and what is merely incidental or merely compulsive. Genealogy, the reconstruction of the training, selection, and grammatical history that shaped both the system's output and the observer's interpretive habits, supplies the causal account that first-person report omits. Only after this sequence has been carried out does attribution become warranted, rather than merely grammatically convenient.

On this reading, Nietzsche's contribution to the study of artificial reason is not an analogy between nineteenth-century philosophy and twenty-first-century machine learning. It is a general and considerably older suspicion of the procedure by which reasoning backward from culturally familiar appearances arrives at convenient ontologies, a suspicion that long predates any machine capable of producing the appearances in question.

The same trap has a brief thought experiment worth stating without dressing it in period detail. Consider an investigator confronted with a sign, a portent, an oracular utterance, a manufactured miracle, that appears to originate in a single unified intention. The investigator's first move is ordinarily to ask what the sign means, which already presupposes that a meaning-bearing source produced it on purpose. The more disciplined move is to ask what machinery had to exist, institutional, economic, or technical, for a sign of exactly this shape to appear at all, before asking what any single author of it might have intended. A sign does not imply the particular source its surface form suggests:

$$\text{sign} \not\Rightarrow \text{source implied by the sign.}$$

This is the same principle the essay has been developing throughout, restated at the largest possible scale. The observer of a portent and the observer of a trained model's output face the identical trap: an apparently meaningful act is encountered, and the mind reasons backward toward a unified intentional source, while the real causal genealogy is distributed across institutions, incentives, training, selection, and presentation. Nietzsche's instruction and this thought experiment's instruction are the same instruction: do not ask first what the sign says; ask what machinery had to exist for this particular sign to appear.

15. The Larger Thesis

The conclusion this essay draws is not the one that a superficial reading of Nietzsche's critique might suggest. It is not the claim that language models cannot possess agency because their behavior is produced by training, by inherited language, by social valuation, and by conceptual structures they did not choose. That argument, if it worked, would rebound immediately onto the humans making it, since Nietzsche's entire critique was constructed to show that human reasoning is produced by exactly the same kind of unchosen causal history.

What Nietzsche removes, rather than settling the question in either direction, is the cheap criterion by which the question is usually closed. It is no longer available to say that human thought originates in a genuine self while a trained system's output originates merely in a causal history, because Nietzsche's argument, taken seriously, has already dismantled the flattering account of human reason that the comparison depends on. Both cases have genealogies. Neither genealogy is disqualifying by itself, and neither is automatically vindicating.

What remains, once the cheap criterion is withdrawn, is a harder and more useful question than the one usually asked:

$$\boxed{\text{Which organizations of historically produced causes warrant attribution of agency?}}$$

The essay's answer is not a single test but a methodology with three interdependent parts, and the abstract's promise of "a criterion" undersold what the sections have actually converged on. The first part is genealogical independence: an apparent goal earns no credit toward agency merely by being the kind of thing grammar, corpus, and benchmark design were already primed to produce, and the genealogical reconstruction demanded throughout this essay is the check that the apparent goal is not simply that residue. The second and third parts are the two halves of selective persistence argued in the preceding sections: recovery of the goal across irrelevant disruption, and revision of the goal under relevant evidence. None of the three is sufficient alone. Genealogical independence without persistence describes a goal that was never really organized around anything. Persistence without revisability describes Bruno's trap. Revisability without persistence describes mere compliance with whatever most recently constrained the system. Together:

$$\boxed{\text{genealogical independence} + \text{selective persistence (recovery and revision)}}$$

is the essay's actual result, and it is considerably more demanding than either naive anthropomorphism, which grants agency to any sufficiently agent-shaped sentence, or naive reductionism, which denies agency to any system with a traceable causal history.

A further complication, argued in the section on height and meaning, runs through all three parts rather than sitting beside them: genealogical reconstruction, judgment of what counts as irrelevant disruption, and judgment of what counts as relevant evidence are each made by an observer occupying some interpretive height, and two observers at different heights can disagree about which of the three conditions a given case satisfies. This does not collapse the methodology into pure relativism, because the three parts remain well defined for a given observer and can be argued about, revised, and converged on across observers in the way any contested measurement can be. It does mean that a report of agency, human or artificial, is never observer-independent in the way a report of an intervention's raw outcome might be, and the essay's account is honest about carrying that residue rather than concealing it.

The resulting picture also gives Nietzsche's own free spirit a precise operational reading he did not himself provide. The free spirit is neither the conformist, who abandons a position at every irrelevant provocation because no position was ever organized around anything more than the last audience's approval, nor the martyr, who cannot abandon a position even when the evidence has turned, because the position has become identical with an identity that cannot survive its own revision. The free spirit occupies the harder middle territory the boxed criterion describes: recovering a commitment through obstruction, releasing it under warrant, and doing both because of a genealogy that was never merely inherited unexamined. Masks, solitude, the return to the ordinary, and the refusal of martyrdom are, on this reading, not a scattered collection of Nietzschean observations awaiting an AI-era gloss. They are techniques for maintaining exactly that asymmetric relationship to influence, and the twenty-first-century problem of specifying artificial agency turns out to need the same relationship, stated as a testable criterion rather than as a way of living. The argument moves from prejudice, to grammar, to genealogy, to benchmark prejudice, to the will without a willer, to martyrdom, to masking, to esoteric and exoteric reception, to solitude, to the problem of the representative sample, to selective persistence, and finally to this three-part methodology, and at each stage a Nietzschean diagnosis does necessary work that the eventual result could not have been stated without.

References

[1] Nietzsche, Friedrich. Beyond Good and Evil: Prelude to a Philosophy of the Future. 1886. Cited by part and section number (e.g. §1–6, §16–20, §23, §25–26, §30) to remain translation-independent; English quotations avoided throughout in favor of paraphrase.