How Do You Attribute Performance in a Complex Socio-Technical Pipeline?

Who Expects What, and What We Are Talking About

Part of the public discourse has cast 2026 as the year of the verdict. Around fifteen compounds originating from artificial-intelligence platforms are entering pivotal Phase III, and we are told these readouts will finally settle whether AI can, or cannot, discover drugs. The thesis of this text is that the question is doubly ill-posed.

- It is ill-posed in substance, because a pivotal trial measures a molecule and not a process.
- It is ill-posed in form, because it seeks an author ("did the AI discover?") where it ought to seek a distribution ("what share of the performance belongs to which component?").

Drug discovery is not an act; it is a chain. One does not attribute a chain to one of its links.

By *capability*, this text will mean one precise and single thing: the reproducible property of a process producing a measurable advantage within a declared domain of use. Three requirements, therefore:

- Reproducible (not a one-off),
- Measurable (not a narrative),
- And declared as to domain (not general by default).

This definition carries the whole article; each time the word reappears, those three conditions must be heard in it, never more.

It is also necessary to name who carries the expectation, because it is not everyone. R&D leadership knows that a pivotal trial does not arbitrate a method. Seasoned investors reason in risk-adjusted net present value, not in validation narrative. The "moment of truth" is a production of the specialized press, of a few influencers, and of a fraction of the funds, who count clinical programs (on the order of 173, of which around fifteen in Phase III, according to sector censuses that must be taken for what they are: tallies by actors with an interest in the high thesis) and read in this volume a promise about to be kept. It is to this narrative that this text addresses itself, before proposing what to replace it with.

Three Levels of Undecidability, and the One We Defend

When one says that performance cannot be attributed to AI, one may mean three things:

1. Level 1: the data are not, today, available,
2. Level 2: these data would be obtainable, but at a prohibitive cost,
3. Level 3: attribution would be theoretically impossible, whatever the data.

This text holds to levels 1 and 2. It does not maintain that AI's contribution is inherently unattributable: it maintains that current arrangements do not allow it to be isolated, and it will later propose those that would. An impossibility of principle would itself be non-falsifiable, hence indefensible. Wherever this text writes "does not allow," one must hear "not with what we observe in 2026," and not "never."

The Survivor, and the Entanglement of the Signal

A drug's pipeline is a sequence of filters.

- A platform proposes candidates,
- Chemists retain or discard them,
- Medicinal chemistry optimizes them,
- ADME eliminates a share, toxicology another,
- Phase I rules on tolerability,
- Phase II on the first efficacy signal.

The base rate is a blunt reminder: across all indications, the probability of moving from Phase II to Phase III is on the order of 30%, and falls below 20% in neurology (Wong, Siah, and Lo, *Biostatistics*, 2019). A Phase III molecule is not a sample: it is a survivor.

One would be tempted to say that the platform's signal dilutes at each filter. That would be false, because modern pipelines are not sequential but iterative: the platform's output informs a human decision, which prompts a new query, which redirects optimization.

The right notion is therefore not dilution but *entanglement*, which must be defined and not merely suggested. Entanglement, here, denotes non-separability: a component's marginal contribution depends on the realized values of the others, because algorithmic outputs and human decisions condition one another by iteration. Two entangled quantities do not subtract. It is this property, not the evaporation of a signal, that obstructs naive attribution.

It commands another, more technical, in turn. The critical narrative readily invokes survivorship bias as a confounding factor: the Phase III population would be biased by survival, hence non-comparable. But if a platform's intrinsic value is precisely to eliminate bad candidates better upstream, then the over-representation of survivors is not an artifact to correct: it is a *mediating variable*, a link on the causal path of the very effect one is seeking. Conditioning on it would block part of what one wants to measure. The problem is therefore not that survival biases the comparison, but that we do not know, as things stand, whether it is a confounder or a mediator. This ambiguity is the identification obstacle, stated properly.

"Discovered by AI" Names Nothing: A Three-Tier Taxonomy

The narrative presupposes a category: there would exist "AI-discovered drugs," which one could enumerate and whose success rate one would measure. This category does not exist in an operational sense. "Discovered by AI" is a communication term, and it aggregates contributions that share neither the same point of application nor the same causal weight.

These contributions fall into three tiers, and conflating them is the first error.

- At the discovery stage: target identification, target validation,
- At the design stage: de novo molecule generation, lead optimization,
- At the development stage: ADMET prediction, biomarker discovery, patient stratification, trial-protocol optimization.

The repositioning of known compounds cuts across all three. These tiers are not commensurable: a platform that identifies a novel target and a platform that reorders a library of lead series do not do the same job, and their success is not counted in the same currency.

To this is added a dimension the narrative ignores: a platform excellent on kinases may be mediocre elsewhere. The pertinent question is therefore not "does AI have a general capability of discovery," but "does it have a capability within a declared domain of use." This is the substitutability logic one applies to digital twins: a system is worth something only within its stated domain of validity, and extrapolating it beyond that domain is a bias, not a generalization. A capability without a declared domain is not a capability: it is an average.

Within Which Causal Framework, and with What Precautions

To assert that AI "does not directly cause" a better molecule has meaning only within a specified causal framework, and one must resist the temptation to invoke them as synonyms. The potential-outcomes framework, inherited from Rubin, defines a component's contribution as a counterfactual contrast: the quantity of interest with the component, minus that same quantity without it, on a comparable unit.

The directed-graph framework, inherited from Pearl, serves to determine whether that contrast is even identifiable from observed data, by distinguishing confounders, mediators, and collider variables. These two frameworks do not answer the same question: one defines the estimand, the other rules on its identifiability. Articulating them in a real industrial pipeline is an open research subject, not a formality; this text mobilizes them as two complementary instruments, without claiming that their reconciliation is settled.

Within this framework, at least one thing is clear: a component's causal contribution is not its rank in the chain, but its marginal contribution. AI upstream is not AI responsible. The question is never "who acted first," but "what would have been different without this component."

The Phase I Anomaly, Without Settling It, and the Question of Time

The narrative has at its disposal an impressive figure: AI-labeled compounds reportedly clear Phase I at 80-90%, against roughly 52% in traditional history. These orders of magnitude come from the same sector censuses, to be consolidated, and the argument does not depend on them. Phase I tests tolerability, not efficacy. Such a gap admits two explanations equally compatible with the data: better design, or selection of candidates close to a chemical space already known to be safe. Both produce the same figure. As long as one does not separate them, the advantage remains unattributable, in the level-1 sense.

In this perspective, it is useful to distinguish platforms that design candidates from those that improve the pipeline's functioning without modifying the intrinsic properties of the molecules. Consider, by way of illustration, an upstream toxicological filtering platform, such as ToxTwin V4. Its objective is not to produce better molecules, but to reduce the probability that a candidate with an unfavorable toxicological profile continues its development. In other words, it does not turn a mediocre molecule into a good one; it modifies the distribution of molecules actually committed to the experimental pipeline.

Such a system can therefore raise the Phase I clearance probability without the retained molecules intrinsically possessing better pharmacological properties. The gain comes from an improvement of the selection function, not of the design function. Experimental resources are concentrated on candidates whose estimated toxicological profile is more favorable, which reduces evaluation costs, lowers early attrition, and improves the pipeline's economic performance, without the algorithm having, on its own, "discovered" a better molecule.

This distinction matters, because it shows that an improvement in the Phase I success rate does not, on its own, allow one to infer a better molecular-design capability. Two causally distinct mechanisms can produce the same observable result: a platform can generate better candidates, or simply select more efficiently those worth evaluating. A clinical readout does not, in itself, allow these two hypotheses to be told apart.

The argument becomes stronger still in an agentic architecture. A molecule generator, a toxicological filtering system, ADMET optimization modules, and a decision orchestrator then form a composite pipeline, in which each component modifies the distribution of candidates passed to the next. Final performance is no longer the property of an isolated algorithm, but that of their composition. Seeking to attribute the success of a clinical candidate to a single component then amounts to confusing the performance of a system with that of one of its elements. What increases is not necessarily the intrinsic quality of the molecules produced, but the collective capacity of the pipeline to transform experimental resources into viable candidates.

A second difficulty compounds the first: the object is not stationary. A platform learns, its teams learn, its datasets and models evolve, its protocols too. The capability of 2024 is not that of 2026, which will not be that of 2032. Asking "can AI discover drugs" without dating the question amounts to measuring a moving target with a fixed ruler. The quantity does not merely have an uncertain value: it has a derivative.

The Symmetry the Trial Forgets

The validation narrative is a poorly instructed trial, because it believes the verdict can fall both ways: a success would amount to acquittal of the method, a failure to its conviction. Both inferences run into the same fact.

Phase III attrition did not wait for AI. For thirty years, molecules designed by the best human chemists have cleared Phase I and Phase II before failing in Phase III, most often on efficacy. Recursion discontinued REC-994 in May 2025 after non-confirmation of trends observed earlier; Insilico published a positive Phase IIa for rentosertib, a TNIK inhibitor, in the GENESIS-IPF trial (71 patients, 22 sites in China): +98.4 mL of forced vital capacity at 60 mg per day against a 20.3 mL decline under placebo (*Nature Medicine*, 2025), before proceeding toward a pivotal Phase III.

The two trajectories coexist under the same label. If the failure of a human molecule never refuted medicinal chemistry, the failure of an AI molecule does not refute the method; and an isolated success validates it no more. A trial bears on a case; to make it the judge of a class is to confuse the outcome with the property.

Let us say it without excessive Popperian rigor: science does not proceed by absolute proofs, but by bodies of evidence, convergence, and plausibility. The reproach addressed to the 2026 readout is not that it lacks a definitive proof, which nothing requires. It is more precise: in the current state of the arrangement, the evidence it provides does not discriminate the hypotheses one claims to have it adjudicate.

Capability Is Not P(success): eNPV at the Center

The debate reduces a platform's capability to a probability of clinical success. This is the central economic error, and correcting it presupposes the right vocabulary. The quantity a pharmaceutical leadership maximizes is not P(success), but the risk-adjusted expected net present value (eNPV): the expectation of a program's future cash flows, weighted by transition probabilities and discounted for time. Three levers enter it, not one. The probability of success, certainly. But also cost, which directly erodes value. And time, which erodes it through discounting: eighteen months saved in research shift cash flows toward the present and increase eNPV, at unchanged probability.

The consequence is that a platform can transform a portfolio's economics without touching biology. Discovering 20% faster and 50% cheaper, even at the price of a few additional points of failure, can raise aggregate eNPV. Serious investors are not, in fact, looking for a better AI: they are looking for a better portfolio, and a better portfolio is built as much by cost structure and iteration speed as by terminal rate. To reduce capability to the discovery of effective molecules is to ignore that the essential value of a platform may reside in what it does to eNPV, not in the efficacy readout.

What Should Be Measured, and in What Unit

The correct question is causal: what is the effect of the process on eNPV, and not "does this molecule work." A unit is still needed. Without a quantity to attribute, a theory of attribution remains philosophy. We retain one, principal: a component's marginal contribution to eNPV, itself decomposable into three interpretable effects, on transition probability, on cost incurred, on time. As a variant, when reasoning upstream of the costing, an informational unit will do: the reduction of uncertainty a component brings about the distribution of viable candidates, measurable as a fall in entropy. The choice of unit is not cosmetic: it fixes what "to contribute" means, and conditions every estimator downstream.

The ideal arrangement would then be a matched arm: same target, same indication, same window, an AI-assisted pipeline against a classical pipeline, compared on chemical diversity, distance to known scaffolds, passage rates, attrition, time, cost. This is an ablation: one removes the component and observes the variation in the chosen unit. One must admit at once that it is unexecutable: no one will fund two competing programs at several hundred million for the sole pleasure of the inference. The realistic path runs through quasi-experiments, familiar to econometrics: historical matching on comparable targets and indications, propensity score, instrumental variables, synthetic controls, target trial emulation, Bayesian causal inference. None is worth a randomized trial; their convergence would form the body of evidence that the isolated readout does not provide.

It will be objected that the patent already offers an objectifiable causal trace: a novel scaffold, its distance to prior art, its inventive character are documented and opposable. This is correct, and useful. But the patent attests the novelty of a structure, not the contribution of a component to performance. A molecule can be patentable and its therapeutic advantage nil; AI may have produced the novelty without producing the efficacy. The trace exists, but it answers a different question.

A Worked Case, Plainly

Take two programs matched on the same target and the same indication, one run without a platform (pipeline A), the other with (pipeline B). Pipeline B reaches Phase III and clears it. What can be concluded? That molecule B works: its performance is established. Nothing more. One cannot conclude that the platform is the cause, because one does not observe the counterfactual (what B would have given without the platform), because human selection may, at each step, have corrected or even produced the advantage attributed to AI, and because A and B are almost never matched in reality.

How would attribution operate? By ablation, one would remove the platform at a precise point of pipeline B, for example lead optimization, replaying that step by the classical route, and one would measure the eNPV gap imputable to that single removal. By marginal contribution of the Shapley type, one would go further: one would compute each component's contribution by averaging its effect over all possible orders of entry into the coalition, so that the sum of contributions reconstitutes the total eNPV. The ToxTwin terrain already practices this discipline at small scale: predicting a toxicity there amounts to estimating under priors in a constrained space, and the system's value is judged there by ablation of its components, never by a favorable case. The same logic holds, on a larger scale, for a discovery pipeline.

Why Shapley, and Under What Conditions

Citing Shapley is not enough; one must say why this one, and where it falters. The appeal of the Shapley value rests on two properties no intuition replaces: efficiency (the sum of contributions equals exactly the total performance, with no unexplained remainder) and symmetry (two interchangeable components receive the same share). These are precisely the guarantees one wants when contributions are entangled and no naive decomposition holds. Other tools exist and do not aim at the same thing: Sobol indices and functional ANOVA decompose the variance of an output, hence answer a question of sensitivity, not of allocation of a scalar value; integrated gradients attribute a model's prediction to its inputs, at the scale of a model, not of a socio-technical pipeline. Shapley is the right candidate when the question is "how to fairly share a total performance among interacting components."

There remains the objection, which is serious: Shapley requires a well-defined coalition, a value function, and a combinatorial cost that explodes with the number of components. In a real pharmaceutical pipeline, none of these conditions is given in advance; defining them is precisely the work. The effective construction of this value function, the operational segmentation of the components, and the calibration of the protocol on real programs belong to an implementation know-how this text does not detail. This is why it speaks of a sketch, and not a method: the Shapley value indicates the form a rigorous attribution would take, it does not exempt one from the work. The tool is sound; its application is a labor of its own.

Performance, Provenance, Capability: The Generalization

The debate persists because it conflates three questions. Performance asks: does this molecule work. Provenance asks: how did it come about, and what share belongs to AI. Capability, in the sense fixed at the outset, asks: is the advantage a reproducible property of the process, within a declared domain, at a known cost. An approval answers the first and stays silent on the third. This trilogy is not a novelty ex nihilo: it extends, within the Institute's corpus, the analysis of the provenance of performance already conducted elsewhere, of which it is the extension into the register of reproducible capability.

It will be objected one last time that pharma is judged by its approvals, and that the law of large numbers dispenses with attributing case by case: if AI compounds get approved en masse more often and more cheaply, that aggregate gap is the capability. The objection would be admissible if the AI-labeled population were comparable to the reference population. It is not: the label aggregates three disjoint tiers, survival filters before the count and plays now as confounder now as mediator, and no matching of target or period corrects these biases. Mass statistics rescue the inference only on homogeneous populations.

This question is not new, and it would be dishonest to present it as an isolated intuition: causal mediation analysis, explainable AI at the level of the system rather than the model, and the algorithmic-accountability literature already pose, elsewhere, the allocation of a result among human and technical components. The present text does not invent it; it imports it into drug discovery, where the media narrative had covered it with a false simplicity.

One must then state one's own conditions of refutation, failing which the critique would itself be unfalsifiable. Here is mine. If, across several hundred cases, independent quasi-experiments converged to show robust reductions in time and cost, superior chemical diversity, and better approval rates at matched target and indication, and if these effects held across several confounder specifications, then the capability hypothesis would become reasonably credible, even without a randomized arm. The convergence of quasi-experiments would stand in for the randomization one cannot afford. As long as that body of evidence does not exist, capability remains to be estimated, not to be proclaimed.

The scope extends beyond the drug, but its domain must be bounded. In any system where production results from a chain of heterogeneous interacting agents, human and algorithmic, the pertinent question is not who produced a result, but how to quantitatively allocate each component's marginal contribution to final performance. Chemistry is one case. Protein generation, laboratory robotics, materials design, industrial optimization satisfy this condition and pose the same question in other guises. Where it is met, attributing a collective result to a single agent is a category error, and allocating it is a measurement program.

There remains the survivor, where we began. It must be returned to its proper function. A compound that reaches Phase III is an excellent indicator of clinical performance: it has survived everything that kills molecules, and that deserves respect. But it is a poor estimator of provenance: its very survival has erased the trace of what produced it. An approval demonstrates that a drug works. It does not demonstrate that the process that produced it constitutes a new industrial capability. 2026 will not tell whether AI can discover drugs. It will tell whether this or that survivor holds, and it will offer, to whoever cares to build it, the first occasion to attribute the rest.

Limits of This Argument

Six reservations, so as not to reproduce the error one denounces:

1. The orders of magnitude on pipeline volume and the Phase I rate come from non-consolidated sector censuses, produced by actors with an interest in the high thesis; they are mobilized as evidence, not as proof, and the argument does not depend on them.

2. The undecidability defended is of level 1-2, conditional on arrangements: a body of converging quasi-experiments would lift it, under the conditions of refutation stated above.
3. The attribution framework is a sketch, not a validated method: the value function, the segmentation into components, and Shapley's combinatorial cost remain to be built programmatically.
4. The Rubin-Pearl articulation is mobilized as complementarity, whereas it constitutes a research problem in its own right.
5. The confounder / mediator distinction applied to survival presupposes a causal graph that remains to be specified case by case.
6. Finally, this text does not address the valuation of these assets (licenses, portfolio, patent claims in the strategic sense): these questions belong to specialized advisory work and exceed the doctrinal perimeter assumed here.

References

1. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273-286. doi:10.1093/biostatistics/kxx069.
2. Xu Z, Ren F, Wang P, Cao J, Tan C, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. Nat Med. 2025;31(8):2602-2610. doi:10.1038/s41591-025-03743-2.
3. Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, editors. Contributions to the Theory of Games II. Princeton (NJ): Princeton University Press; 1953. p. 307-317.
4. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688-701.
5. Pearl J. Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge: Cambridge University Press; 2009.