Self-Evolution Is Not Enough: Toward a Historically Materialist AI for Scientific Problem Solving

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

Recent work on self-improving agents, evolutionary search, autonomous scientific discovery, and open-ended artificial intelligence has substantially expanded the range of processes through which an artificial system can revise hypotheses, programs, strategies, tools, and even parts of its own architecture. Yet self-improvement does not by itself amount to historical self-evolution. Scientific problems do not exist in an abstract and temporally uniform solution space. What can be formulated, investigated, tested, and established at a given moment depends upon historically accumulated concepts, evidence, instruments, methods, forms of labor, institutions, infrastructures, and relations of knowledge production. This paper develops a philosophical account of artificial intelligence for scientific problem solving from this observation. Drawing on Hegelian dialectics, historical materialism, and an analysis of the unequal structural roles of contradictions within developing systems, it treats problem solving as the transformation of a historically constituted problem situation rather than as the optimization of candidate answers under a fixed objective. The central proposal is that a genuinely self-evolutionary scientific AI would need to reconstruct the material and relational conditions under which a problem has become possible, identify contradictions that constrain or organize further development, distinguish structurally central contradictions from derivative ones, compare interventions by the research states they make reachable, and revise its own representation of the problem after those interventions alter the conditions of inquiry. On this view, an apparently minor intermediate problem may be scientifically important because solving it creates data, methods, instruments, distinctions, institutions, or coordination capacities through which a previously inaccessible problem becomes researchable. The paper therefore distinguishes conceptual possibility, present researchability, and epistemic establishability, and introduces the problem of epistemic stage-skipping in AI-assisted research. It further argues that methods and evaluative regimes are themselves historically situated components of scientific production. A method or criterion that once enables inquiry can become restrictive as the material, conceptual, and institutional capacities of research develop. Under such conditions, stronger optimization against an inherited evaluator can deepen epistemic misalignment rather than resolve it. Historically self-evolutionary AI must therefore be able to diagnose when methods or evaluators have exceeded their domain of adequacy and to transform them through a documented lineage that preserves what remains valid. The resulting account is neither an engineering architecture nor a claim that Hegelian or Marxian concepts can be directly converted into algorithms. It is a philosophical design orientation for long-horizon artificial intelligence whose task is not merely to produce better answers, but to participate in the historical transformation of what can responsibly be asked, known, tested, and solved.

Keywords: artificial intelligence; self-evolutionary AI; dialectics; historical materialism; scientific problem solving; autonomous science; contradiction; historical development; philosophy of science

Discussion Paper Note

This paper is a philosophical and exploratory inquiry into the conditions under which artificial intelligence could participate in long-horizon scientific problem solving as a historically self-evolutionary process. Its point of departure is a distinction that becomes increasingly important as artificial systems acquire capacities for iterative hypothesis generation, program evolution, autonomous experimentation, self-modification, and open-ended search. An artificial system can improve itself, refine a candidate solution, accumulate successful strategies, or preserve evolutionary stepping stones without yet possessing an adequate account of the historical development of the problem situation in which those activities occur. The present inquiry therefore treats self-evolution as a problem concerning the evolution of scientific solvability, rather than only the evolution of an agent or the quality of its outputs.

Contemporary work already provides substantial reasons to avoid a weak formulation of this problem. Evolutionary and self-improving systems can generate variants, evaluate them, retain successful candidates, modify their own tools or code, preserve diverse lineages, and recursively use earlier products as resources for later search. AlphaEvolve combines large language models with automated evaluators and an evolutionary program database in order to develop and improve algorithmic solutions [7]. The Darwin Gödel Machine modifies its own agent code, evaluates those modifications empirically, and preserves an archive of agents that can serve as stepping stones for later development [26]. The AI Co-Scientist organizes specialized generation, reflection, ranking, evolution, proximity, and meta-review processes around a scientist-specified research goal [8]. The AI Scientist-v2 carries hypothesis formulation, experimental execution, analysis, visualization, and manuscript production through an agentic tree-search process [24]. These systems make it increasingly implausible to identify the frontier merely with iteration, recursive improvement, evolutionary variation, or autonomous research workflow.

The philosophical problem considered here begins one level outward. Scientific inquiry takes place within historically accumulated conditions that determine which questions can be formulated with sufficient precision, which hypotheses can be tested, which observations can be obtained, which distinctions can be operationalized, which experiments can be afforded, which results can be reproduced, and which claims can be responsibly established. These conditions include conceptual knowledge and formal methods, but they also include instruments, datasets, laboratories, computation, skilled labor, institutions, standards, property and access relations, infrastructures of communication, systems of coordination, and inherited patterns of disciplinary organization. A problem may therefore be logically conceivable while remaining scientifically inaccessible. Conversely, an apparently modest problem may be historically decisive because its solution changes the conditions under which later and more difficult problems can be investigated.

The paper develops this claim through a selective engagement with Hegelian dialectics, historical materialism, and a hierarchy-sensitive analysis of contradictions within developing systems. These traditions are not treated as a complete computational theory, nor are their concepts assumed to admit straightforward translation into algorithms. They instead provide philosophical resources for identifying questions that current accounts of self-evolving artificial intelligence leave underdeveloped. Hegelian dialectics directs attention to internally generated limitation, contradiction, mediation, and transformation. Historical materialism directs attention to the material and social conditions through which a form of activity becomes possible and through which its contradictions acquire determinate force. A hierarchy-sensitive analysis of contradictions directs attention to the unequal structural importance of tensions within a complex process and to the possibility that this ordering changes as the process develops.

The Hegelian contribution is especially important because the paper rejects a superficial identification of dialectics with the familiar formula of thesis, antithesis, and synthesis. A computational system can easily generate an opposing proposition to a given claim and then compose a verbal compromise. Such behavior may be useful for deliberation, but it does not by itself reproduce the philosophical problem of immanent development. In Hegel's dialectical account, a determination becomes unstable through its own one-sidedness or restrictedness; its transition is generated from within the content rather than supplied arbitrarily from outside, and the later determination both cancels and preserves aspects of what

DRAFT

preceded it through sublation, or *Aufhebung* [17]. For artificial intelligence, the corresponding challenge is therefore not simply to produce a counterposition. It is to identify a limitation that arises from the structure of the current problem representation itself, determine how that limitation constrains further inquiry, and construct a transformation in which relevant achievements of the earlier state remain available within a reorganized problem situation.

This requirement changes the meaning of contradiction. A contradiction in scientific development need not be a pair of formally inconsistent propositions. It can be a relational incompatibility among demands, capacities, methods, institutions, or material conditions that jointly structure a research process. A field may demand increasingly realistic models while lacking sufficiently informative data; require extensive replication while experiments become increasingly costly; possess rapidly expanding data while interpretive theory develops more slowly; or require interdisciplinary integration while incentives, training, and publication systems reproduce disciplinary separation. Such tensions are historically situated. Their importance depends upon what other capacities are available, which institutions organize the field, and how previous interventions have transformed the research environment.

Historical materialism adds a second requirement. Marx's 1859 preface to *A Contribution to the Critique of Political Economy* emphasizes the relation between developing productive forces and historically specific relations of production, and it links the emergence of solvable historical tasks to the maturation of material conditions for their solution [16]. The present paper does not equate scientific inquiry with economic production and does not assume that concepts developed for political economy can be transferred without modification. It takes from this tradition a more abstract question: what capacities actually produce inquiry, and what relations organize access to, coordination of, and control over those capacities?

For scientific problem solving, the productive forces of inquiry may include human expertise, artificial intelligence, formal methods, datasets, measurement instruments, experimental platforms, laboratories, computation, software, accumulated scientific knowledge, and capacities for observation and intervention. Relations of knowledge production may include institutional authority, division of epistemic labor, ownership, funding, data access, publication systems, disciplinary boundaries, coordination structures, evaluation procedures, and relations among researchers, infrastructures, and communities. These categories are provisional analytical devices rather than direct equivalents of Marxian economic categories. Their purpose is to prevent an artificial research system from treating every scientific obstacle as an informational or cognitive deficit.

This distinction matters because a field can possess considerable theoretical and technical capacity while remaining constrained by the relations through which that capacity is organized. Data can exist without being interoperable or accessible. Instruments can exist without sufficient coordination for replication. Expertise can exist in separated communities whose institutional incentives discourage sustained collaboration. Computational capacity can expand while access becomes concentrated. A method can be technically available while norms of evaluation prevent its results from entering a field's accepted evidential practices. Under such conditions, generating a more sophisticated hypothesis may have little effect on the underlying obstacle. A historically materialist artificial intelligence would need to distinguish a conceptual bottleneck from a material bottleneck, an evidential bottleneck from an institutional bottleneck, and a lack of productive capacity from a relation that prevents existing capacity from being effectively combined.

A third philosophical question concerns the unequal structural significance of contradictions within a complex process. When a research situation contains many tensions, which one currently has the greatest role in reproducing, constraining, or organizing the others? The paper treats this as a relational and temporally variable question rather than as a permanent label attached to a problem. A contradiction may be structurally central at one stage because several other bottlenecks depend upon it, yet lose that role after an intervention changes the research environment.

A future system might therefore represent a research situation at time t by a state S_t containing currently relevant knowledge, evidence, methods, instruments, infrastructures, institutional conditions, access relations, and accumulated historical constraints. From this state there emerges a set of presently

DRAFT

meaningful and reachable research questions, denoted provisionally by $\mathcal{Q}(S_t)$, together with a set of contradictions or developmental tensions $\mathcal{C}(S_t)$. A structurally central contradiction, if such a representation is useful for a particular case, should not be defined merely as the contradiction whose description sounds most important. Its status would depend upon relations of dependency, bottleneck effects, persistence, reproduction, causal leverage, and influence upon the future reachability of inquiry. An intervention that changes a structurally central contradiction may reorganize several dependent tensions at once, while an intervention directed at a vivid but derivative tension may yield only local improvement.

The hierarchy of contradictions must itself remain historically revisable. If a new instrument resolves a measurement bottleneck, the previously central contradiction may lose its structural role and another tension may become decisive. A field constrained by scarce data can become a field constrained by interpretability after measurement capacity expands. A field constrained by computational cost can become a field constrained by validation after computation becomes abundant. A system capable of dialectical historical reasoning must therefore reconstruct the contradiction structure after intervention rather than assume that the original problem decomposition remains valid throughout the research process.

This leads to the paper's central distinction between candidate evolution and problem evolution. Candidate-centered evolutionary systems take a task, evaluator, environment, or research goal as sufficiently stable for a search process to improve programs, hypotheses, strategies, or agents. More open-ended systems relax some of these assumptions by preserving diverse stepping stones or by allowing the effective environment to change through co-evolution. The present argument accepts the significance of these developments. It nevertheless asks whether the problem-defining conditions themselves are represented as objects of historical transformation. The relevant transition is from optimizing x under a problem representation to transforming the state S_t that determines which x can count as a meaningful candidate and which criteria can count as adequate evaluation.

A provisional expression of this difference is

$$S_t \xrightarrow{a_t} S_{t+1},$$

where a_t is a research intervention rather than merely a candidate answer. The intervention can be theoretical, empirical, methodological, infrastructural, institutional, or coordinative. It can create a dataset, introduce a conceptual distinction, build an instrument, falsify an attractive hypothesis, establish a negative result, standardize a protocol, reorganize access to evidence, or connect previously separated research communities. What matters is that the action changes the future research state and thereby changes the set of subsequently reachable questions.

This perspective gives intermediate problems a role that ordinary decomposition does not fully capture. A difficult problem P can be inaccessible from S_t even when an artificial system can linguistically formulate a plausible answer to it. The appropriate next action may therefore concern a smaller problem p_1 whose solution does not approximate P in any simple metric. Solving p_1 may instead produce a new method, distinction, dataset, or instrument such that the transition $S_t \to S_{t+1}$ makes another problem p_2 investigable. A sequence of such transformations can eventually place P within the reachable research space. The scientific significance of p_1 then lies partly in its generative effect upon future inquiry.

This is one reason the paper distinguishes conceptual possibility, researchability, and epistemic establishability. Conceptual possibility concerns whether a proposition, mechanism, or explanatory structure can be coherently formulated. Researchability concerns whether available methods and conditions permit disciplined inquiry into it. Establishability concerns whether the available evidential and methodological relations permit a claim to acquire an appropriate degree of epistemic warrant. Artificial intelligence can collapse these levels because it can synthesize conceptual resources drawn from many historical periods and disciplinary contexts at once. A model may therefore construct a sophisticated theoretical state whose intermediate empirical, methodological, and institutional conditions have never actually been developed.

The paper provisionally calls this problem *epistemic stage-skipping*. Stage-skipping does not mean that every research programme must pass through a predetermined sequence or reproduce the historical order through which an earlier field developed. Scientific innovation often results precisely from recombination, conceptual shortcuts, new instruments, and unexpected crossings of established bound-

DRAFT

aries. The problem arises when a system produces the appearance of a later-stage theoretical achievement while the conditions required to investigate, discriminate, validate, or reproduce its claims remain absent. The danger is therefore epistemic rather than chronological. The relevant question is whether a proposed transition has generated or otherwise secured the conditions required for responsible establishment of the resulting claims.

A historically self-evolutionary system would consequently need some representation of reachability. It should be able to distinguish a problem that is merely difficult from one whose current conditions of investigation are incomplete. It should ask which missing condition is decisive, which intermediate action can alter that condition, and how alternative actions would reshape later possibilities. The system may therefore sometimes respond to a grand question by recommending that the grand question remain temporarily unresolved while a seemingly narrower research task is undertaken. Such restraint would not indicate failure of intelligence. It could express a more developed understanding of how scientific intelligence participates in the production of future solvability.

Methods and evaluators become part of this problem. Many successful evolutionary systems rely upon an objective or at least operational evaluator. AlphaEvolve is especially effective where proposed programs can be systematically executed and scored [7]. Scientific development frequently includes domains in which both methods and criteria of evaluation evolve. New instruments change what counts as measurable. New statistical methods alter standards of evidential adequacy. New conceptual distinctions reveal that an older benchmark conflated different phenomena. Replication crises can transform methodological expectations. Changes in disciplinary organization can modify what kinds of evidence are considered admissible or salient. A method or evaluator that once enabled cumulative inquiry can therefore become a constraint after the material and epistemic state has changed. A long-horizon artificial research system must allow that the method and function by which progress is produced and evaluated at t may themselves become objects of criticism and reconstruction at $t + 1$.

This does not imply that evaluation should become arbitrary or that every historical change invalidates prior standards. The issue is that criteria require provenance and scope. A criterion can be adequate for one research state and become insufficient after the object, available evidence, or method changes. Dialectical development therefore includes the possibility that an evaluator contains an internal limitation: by successfully optimizing against it, a system can expose properties that the evaluator failed to represent. A historically self-evolutionary AI should be capable of recognizing such cases and distinguishing genuine progress from Goodhart-like adaptation to an obsolete or partial criterion.

Historical memory is similarly broader than accumulation of previous outputs. An archive of successful and unsuccessful candidates can be extremely valuable, as open-ended evolutionary systems already demonstrate. The present account asks for additional relational structure. A historically meaningful memory should preserve why an intervention was attempted, what material and conceptual conditions were present, what contradiction it addressed, which alternatives were excluded or unavailable, what unexpected effects followed, and which possibilities were opened or closed. A failed experiment can then function as more than a negative score. It can alter the representation of the problem, expose an incorrect causal assumption, eliminate a developmental path, or reveal that a previously secondary contradiction has become decisive.

Such memory also guards against a teleological account of self-evolution. The paper does not assume that scientific history proceeds through a single sequence of progressively superior states. Historical processes branch. They contain lock-in, regression, forgotten alternatives, inaccessible paths, rediscovered concepts, institutional inertia, and transformations that enlarge some capacities while reducing others. The transition from S_t to S_{t+1} can therefore open and close possibilities simultaneously. A historically self-evolutionary system should preserve plural developmental trajectories where uncertainty remains material rather than prematurely compress them into one winning lineage.

Counterfactual historical reasoning becomes useful here. When comparing interventions, an AI can ask not only which action appears most promising according to the current evaluator, but how different actions would transform the future research landscape. An intervention that produces a small immediate result may create a widely reusable dataset. Another may produce a striking result while increasing

DRAFT

dependence upon an inaccessible infrastructure. A third may reveal that two apparently competing theories concern different regimes and thereby dissolve a false opposition. Their values cannot be compared solely through the magnitude of the next output. They differ in the research states they make reachable and in the dependencies they create.

The term *generative value* is used provisionally for this future-oriented dimension of research significance. Generative value concerns the capacity of a research action to alter what can subsequently be asked, observed, tested, combined, or established. It is not proposed as a single scalar measure, and the paper does not assume that a universal objective function could aggregate all such effects. The concept instead directs attention to consequences that immediate-performance metrics can miss. Datasets, protocols, null results, conceptual distinctions, shared infrastructures, replications, and negative theoretical results may have modest direct novelty while possessing substantial generative significance for later inquiry.

The resulting account also changes the meaning of autonomous problem solving. The familiar representation $P \to A$ treats the problem as given and asks for an answer. A historically self-evolutionary representation is recursive. The system begins from a historically constituted state, reconstructs the relevant contradictions, selects or proposes an intervention, observes how the intervention changes the material and epistemic situation, and then formulates the problem again. Schematically,

$$S_t \to \mathcal{C}_t \to C_t^* \to a_t \to S_{t+1} \to \mathcal{C}_{t+1} \to \cdots,$$

where C_t^* denotes a contradiction treated provisionally as structurally central at time t. Every component remains revisable. The system can discover that its initial contradiction map was mistaken, that the proposed structurally central contradiction was derivative, that an intervention generated unanticipated dependencies, or that the criteria used to evaluate the transition were themselves incomplete.

This recursive structure is also where dialectics becomes more than a method of argument generation. A system that generates a thesis, invents an antithesis, and produces a synthesis can remain entirely within a fixed problem representation. A dialectically oriented historical system would instead ask whether the current determination generates a limitation through its own operation, whether the opposition reflects a real dependency or material tension, what aspects of the existing state should be preserved, and whether a proposed mediation actually changes the relation that reproduced the contradiction. The resulting state must then be examined for new contradictions rather than treated as a final synthesis.

The paper therefore adopts a non-finalist understanding of solution. Some contradictions can and should be eliminated, especially when they arise from error, avoidable deprivation, or unnecessary institutional obstruction. Other tensions accompany the development of a research system and reappear in transformed forms. Increasing measurement resolution can create new problems of interpretation. Increasing automation can create new problems of validation, provenance, and control. Increasing interdisciplinary integration can create new coordination and translation problems. The aim is consequently not contradiction for its own sake, nor perpetual instability, but an ability to understand which transformations enlarge responsible inquiry and which merely reproduce or displace existing obstacles.

The concept of productive forces and relations of production introduces a further caution. An AI that diagnoses scientific development through historical material categories can easily reproduce reductive analogies. Scientific inquiry contains normative, interpretive, conceptual, and experiential dimensions that cannot be exhausted by an inventory of material resources. Conversely, an account focused only on concepts can miss the infrastructures and relations through which those concepts acquire practical research force. The paper therefore treats material and conceptual conditions as mutually constraining elements of a historically generated research state. It seeks neither an economic reduction of epistemology nor an idealist account in which better reasoning alone dissolves material obstacles.

A second caution concerns the identification of structurally central contradictions. Any system capable of ranking contradictions can produce severe errors if it mistakes salience for structural importance, inherits ideological assumptions from its training data, or models a field through an incomplete set of actors and relations. The designation of a contradiction as structurally central can itself influence resource allocation and thereby become performative. A research AI that declares one bottleneck central may

DRAFT

redirect experiments, funding, attention, and institutional coordination toward it, potentially suppressing alternatives. Such inference must therefore remain contestable, provenance-sensitive, uncertainty-aware, and open to competing models of the same research situation.

A third caution concerns historical materialism itself. The paper uses the tradition as a philosophical foundation for asking how material conditions, social relations, contradiction, and developmental possibility interact. It does not assume that all historical change follows one necessary sequence, that productive development automatically yields epistemic progress, or that a scientifically more capable state is normatively superior in every respect. An artificial system should therefore avoid converting historical materialism into a hidden teleological objective function. Development must remain distinguishable from moral improvement, social justice, or human flourishing, each of which requires additional normative argument.

A fourth caution concerns the autonomy of scientific AI. Historical self-evolution does not imply that an artificial system should receive unrestricted authority to transform institutions, infrastructures, or research agendas. The capacity to model historical conditions and compare interventions is distinct from legitimate authority to implement them. Many interventions involve allocation of resources, access to communities, experimentation on human or nonhuman subjects, environmental effects, institutional governance, and distributive consequences. The present paper concerns the philosophical capacities required for historical problem solving; it leaves questions of authorization, accountability, democratic governance, and limits of autonomous intervention for separate normative treatment.

The paper's positive proposal is therefore intentionally modest. A future historically self-evolutionary AI would require, at minimum, capacities for reconstructing historical research states, distinguishing conceptual from material and institutional constraints, identifying immanent tensions, representing dependence among contradictions, assessing which contradictions are structurally central under current conditions, comparing interventions by the future research states they make reachable, preserving developmental alternatives, revising evaluative criteria when their limitations become manifest, and reformulating the problem after intervention. These are functional and philosophical requirements rather than commitments to a particular neural architecture, agent framework, symbolic representation, or optimization algorithm.

Several of these capacities may ultimately require hybrid methods. Historical reconstruction may combine language models, causal models, scientific databases, provenance systems, institutional data, and human judgment. Contradiction mapping may require representations of dependencies and constraints that go beyond textual similarity. Reachability analysis may require simulations, planning systems, uncertainty models, and explicit representations of missing evidence or missing infrastructure. Reevaluation may require meta-scientific models capable of distinguishing failure of a candidate from failure of the criterion used to judge it. The paper leaves these engineering questions open because its purpose is prior: to clarify what kind of evolution is being requested before specifying how to build it.

This prior clarification is important for the language of self-evolution. An artificial system can evolve its own code while continuing to optimize within a historically frozen representation of the world. It can accumulate scientific papers while treating the institutions and instruments that made those papers possible as invisible background. It can generate diverse hypotheses while assuming that the research goal and evidence regime have already been correctly specified. It can preserve evolutionary stepping stones while lacking a concept of why one stepping stone transformed the conditions of later inquiry. Such systems may be powerful forms of self-improvement. The present paper reserves the stronger idea of historical self-evolution for systems that participate in the transformation and reconstruction of the problem-producing conditions themselves.

The resulting thesis can be stated compactly. A genuinely self-evolutionary scientific AI should not merely improve its answers through time. It should model how the historical conditions of inquiry generate a present problem space, identify the contradictions that organize or constrain that space, intervene in ways that transform future reachability, and revise both its problem representation and its criteria of progress as those interventions change the research world. The object of evolution is therefore distributed across the agent, its representations, the scientific environment, and the relations through which knowl-

DRAFT

edge is produced.

This account remains exploratory. The concepts of research state, contradiction hierarchy, historical reachability, generative value, and epistemic stage-skipping require further differentiation and may not admit uniform treatment across disciplines. Physics, mathematics, clinical science, field anthropology, historical research, and social inquiry involve very different relations among theory, evidence, instruments, subjects, institutions, and evaluation. A future theory must therefore preserve domain heterogeneity rather than force every research process into one developmental grammar. The present discussion paper offers a philosophical orientation for that work: self-evolution becomes scientifically significant when artificial intelligence can reason about the history through which solvability itself is produced.

1 Introduction

Artificial intelligence is increasingly being developed not only as a system that answers questions, but as a system that can participate in iterative processes of search, evaluation, revision, and scientific discovery. Recent work illustrates several distinct forms of this development. AlphaEvolve combines large language models with automated evaluators and an evolutionary program database so that candidate programs can be generated, tested, retained, and used as material for subsequent search [7]. The Darwin Gödel Machine goes further by modifying parts of its own agent code, empirically evaluating those modifications, and preserving a growing archive of agents that can serve as later evolutionary stepping stones [26]. The AI Co-Scientist organizes multiple specialized agents around generation, reflection, ranking, evolution, and meta-review in order to iteratively improve scientific hypotheses relative to a researcher-specified goal [8]. The AI Scientist-v2 extends autonomous scientific workflow through agentic tree search that connects hypothesis formation, experiment design and execution, analysis, visualization, and manuscript production [24]. Taken together, these systems make clear that iteration, recursive improvement, evolutionary variation, branching search, and partial automation of scientific practice are no longer merely speculative features of future artificial intelligence.

These developments nevertheless leave open a different question about what it means for artificial intelligence to become genuinely self-evolutionary in scientific problem solving. A system may improve its own code, generate increasingly successful hypotheses, search through alternative programs, preserve useful predecessors, or repeatedly optimize experimental choices while the problem situation in which these activities occur remains comparatively fixed. The research goal may already have been specified, the available variables may already have been selected, the relevant evidence may already be assumed obtainable, and the criteria by which progress is evaluated may already be given. Under such conditions, the system can evolve within a problem space without representing how that problem space itself came into being, what historical conditions presently sustain it, which of those conditions constrain further development, or how an intervention might change what becomes scientifically possible afterward. The distinction developed in this paper begins here: self-improvement concerns changes in an artificial agent or its outputs, whereas historical self-evolution concerns changes in the relation among the agent, the problem representation, the material conditions of inquiry, and the historically developing space of solvable problems.

Scientific inquiry provides strong reasons to take this distinction seriously. Many difficult problems cannot be solved merely by applying greater reasoning capacity to their final formulation. Their eventual solvability depends upon a long development through intermediate concepts, measurements, instruments, datasets, mathematical techniques, experimental practices, institutions, and forms of coordination. A research question can be intelligible before it is empirically investigable, and empirically investigable before its answer can be responsibly established. Conversely, work that appears narrow or preliminary when judged by its immediate conclusion can be historically decisive if it produces a dataset, method, distinction, protocol, instrument, negative result, or institutional capacity that transforms later inquiry. Scientific development is therefore not adequately represented as a sequence in which a stable problem receives progressively better candidate answers. The conditions under which the problem can be posed, tested, and judged develop together with the inquiry itself. In this sense, the history of research is partly

DRAFT

a history of the production of solvability.

This paper proposes that a stronger conception of self-evolutionary artificial intelligence should take that production of solvability as part of its object. The relevant system is not only an AI model or agent, but a historically constituted research situation that contains established knowledge, available evidence, methods, instruments, infrastructures, skilled labor, institutional arrangements, access relations, standards of evaluation, and accumulated paths of earlier success and failure. An artificial system oriented toward long-horizon scientific development would need to reason about how these heterogeneous conditions jointly make some questions reachable while leaving others premature. It would also need to distinguish obstacles that arise from conceptual insufficiency from those that arise from absent evidence, inadequate instruments, fragmented institutions, restricted access, missing coordination, or incompatible evaluative practices. The point is not to deny the importance of inference, creativity, or hypothesis generation. It is to situate those capacities within the historical and material conditions that determine whether a proposed insight can become scientific knowledge.

Hegelian dialectics provides one philosophical resource for describing this kind of development. The paper does not treat dialectics as a mechanical sequence of thesis, antithesis, and synthesis, nor as an instruction to generate a contrary statement whenever a position is encountered. Such a procedure can produce useful debate while leaving the original problem representation intact. The more relevant Hegelian insight concerns immanent limitation: a determination can generate tensions through its own structure, and its development can require a transformation that both negates and preserves aspects of what preceded it through *Aufhebung* [17]. Translated cautiously into the present problem, an artificial research system would need to ask whether a limitation is external to a research program or generated by the way the program itself organizes concepts, methods, evidence, and aims. A scientifically important transition would then consist not merely in choosing between opposed positions, but in reorganizing the relations that produced the opposition while retaining whatever remains generatively necessary from the earlier state.

Historical materialism adds a second orientation by directing attention from conceptual development to the material and social conditions through which forms of activity become possible. Marx's account of the relation between productive forces, relations of production, and historically emergent tasks cannot simply be transferred into a computational theory of science [16]. Scientific inquiry is not reducible to economic production, and the present paper does not claim an isomorphism between the two. The historical-materialist question is nevertheless valuable: what capacities actually produce a field's knowledge, what relations organize those capacities, and under what conditions do existing relations become enabling structures or constraints on further development? For scientific inquiry, productive capacities may include expertise, instruments, computation, methods, datasets, laboratories, software, and accumulated knowledge, while relations of knowledge production may include ownership, funding, institutional authority, division of epistemic labor, publication systems, disciplinary boundaries, access arrangements, and coordination structures. A historically materialist AI would therefore need to represent not only what is known, but how the capacity to know is materially and institutionally organized.

An important consequence follows for method and evaluation. If methods, measurement practices, evidential standards, and evaluative procedures are part of the organized capacity through which knowledge is produced, they cannot be treated as permanently external to history. A method can be an enabling condition at one stage and later become a constraint when new instruments, data regimes, conceptual distinctions, or institutional arrangements change the object and practice of inquiry. The same applies to evaluators. A criterion developed for one research state can continue to organize selection after the state has changed, creating historical misalignment between what is optimized and what now counts as scientific progress. In that case, optimizing more effectively under the inherited criterion can reproduce or intensify the limitation. A historically self-evolutionary AI therefore needs to reason not only about changing answers and problems, but about the historical adequacy of the methods and evaluative regimes through which answers are produced and judged.

This orientation also changes how contradiction should be understood in artificial problem solving. Contradiction need not refer only to formal inconsistency between propositions. A developing research

DRAFT

process can contain tensions among requirements, resources, methods, and institutional arrangements: increasingly realistic models can demand information that available measurements cannot supply; demands for replication can conflict with rising experimental cost; rapidly expanding datasets can outpace interpretive theory; interdisciplinary problems can require forms of integration that existing disciplinary incentives make difficult to sustain. Several such contradictions can coexist, yet they need not have equal structural importance. One may organize or reproduce several others at a particular stage, while later losing that role after an intervention changes the research state. The corresponding problem for artificial intelligence is therefore hierarchy-sensitive and temporal: how can a system distinguish a structurally central contradiction from derivative tensions, justify that distinction through relations of dependence rather than rhetorical salience, and recognize when the hierarchy has changed?

The answer cannot be a fixed ranking rule applied once to a static description. If a research intervention changes the available evidence, instrumentation, institutions, or relations among participants, then the contradiction structure must be reconstructed. What counted as the principal bottleneck can disappear, fragment, or be displaced by a new limitation generated by the successful intervention itself. Increasing measurement capacity can transform a problem of data scarcity into a problem of interpretation. Automating experiments can transform a problem of labor scarcity into problems of provenance, validation, and control. Building a shared infrastructure can remove a coordination barrier while creating new questions of governance and access. Historical self-evolution therefore requires recursive re-problematization: after acting on the research situation, the AI must ask again what the problem has become. The trajectory of inquiry matters because the later problem is generated through transformations of the earlier state rather than simply selected from a timeless inventory of already defined questions.

This view gives special importance to intermediate research. An apparently easy problem can be scientifically valuable not because it approximates the final answer in a linear decomposition, but because its solution changes the state from which future inquiry proceeds. A standardized measurement protocol can make previously incomparable observations commensurable. A null result can eliminate an entire family of research trajectories. A conceptual distinction can render an earlier question too coarse and thereby generate better questions. A new instrument can convert a theoretical dispute into an empirical one. A shared dataset can enable a field to test hypotheses that were previously discussed only speculatively. These cases suggest a notion of generative research value that differs from immediate performance, novelty, or publication significance. The relevant question is partly prospective: what new research states become reachable because this intervention has occurred? A historically self-evolutionary AI should therefore be capable of recommending work whose principal contribution is to enlarge the future space of responsible inquiry rather than to maximize an immediate score.

The same argument exposes a distinctive risk of AI-assisted scholarship: epistemic stage-skipping. Large language models compress linguistic traces from many periods, disciplines, and stages of scientific development into a common inferential resource. This enables productive cross-domain recombination, but it also makes it easy to construct a theoretically mature-looking answer whose intermediate empirical, methodological, or institutional prerequisites have not been produced. A proposal can be linguistically coherent, conceptually imaginable, and even suggestively correct while remaining scientifically ungrounded because the evidential chain needed to establish it does not yet exist. The danger is not only hallucination in the ordinary sense of false factual generation. It is the generation of an answer appropriate to a later research state while presenting it as though the present state already supplied the necessary conditions of justification. Distinguishing conceptual possibility, present researchability, and epistemic establishability is therefore central to responsible long-horizon scientific AI.

A further consequence is that evaluation itself may need to evolve. Evolutionary and self-improving systems are most straightforward when candidate quality can be tested against a stable evaluator, benchmark, or objective. AlphaEvolve illustrates the power of this arrangement in domains where proposed algorithms can be automatically verified and quantitatively scored [7]. Yet some historical research transitions alter not only candidate solutions but the relevant methods, evidence standards, or problem definitions. In such cases, a fixed evaluator can become part of the obstacle. A historically self-evolutionary system would need to distinguish failure under an adequate criterion from evidence that the criterion

DRAFT

no longer captures what the transformed research situation requires. This does not imply that evaluative standards should be freely rewritten whenever results are inconvenient. On the contrary, revisions would need stronger justification, provenance, contestability, and comparison with preserved alternatives. The philosophical point is narrower: the criteria of scientific progress belong to the developing research situation and cannot always be assumed external to history.

The resulting conception of self-evolution is therefore distributed across several coupled transformations. The artificial agent can change its strategies, tools, and representations; the research problem can be reformulated; material and institutional conditions can be altered by scientific action; contradictions can change in relative structural importance; and evaluative criteria can themselves become objects of scrutiny. None of these changes guarantees progress, and their conjunction should not be interpreted teleologically. Scientific histories contain branching trajectories, dead ends, lock-in, regressions, rediscoveries, suppressed alternatives, and transformations that enlarge some capacities while narrowing others. A historically self-evolutionary AI must therefore preserve uncertainty about developmental direction rather than treating history as a monotonic ascent toward a predetermined optimum. Its task is not to predict or impose an inevitable historical sequence, but to reason about historically situated possibilities, dependencies, and consequences while remaining capable of revising its own developmental model.

The paper develops this proposal as a philosophical design orientation rather than an engineering architecture. Section 2 establishes the conceptual foundations in dialectical development and historical materialism. Section 3 reviews contemporary evolutionary AI, recursive self-improvement, autonomous scientific agents, and related approaches in order to clarify what is already possible and where the remaining conceptual gap lies. The subsequent sections develop the distinction between agent evolution and historical problem evolution, characterize the material and relational conditions of scientific solvability, introduce the idea of a historically constituted research state and a hierarchy of contradictions, and examine the generative role of intermediate problems. The analysis then turns to historically self-evolutionary problem solving, the co-evolution of problems, methods, and evaluative criteria, historical memory and branching development, and the functional requirements that a future AI would need to realize. The final sections address limitations and risks, including pseudo-dialectics, false identification of structurally central contradictions, teleological reduction, institutional reproduction, and historical-model lock-in. The central thesis throughout is that self-evolution becomes historically significant only when artificial intelligence can participate in the transformation and reconstruction of the conditions through which scientific problems become solvable.

2 Philosophical Foundations of Historical Self-Evolution

The argument of this paper depends on a conception of development that is stronger than repeated optimization and broader than the self-modification of an artificial agent. The present section establishes that conception before the contemporary AI literature is reviewed. Its purpose is not to reconstruct Hegel or Marx exhaustively, nor to claim that their philosophical systems can be translated directly into a computational architecture. It isolates a limited set of ideas that are useful for thinking about long-horizon problem solving: development generated from internal limitations, transformation that both negates and preserves earlier structures, the dependence of possible change on materially organized conditions, the unequal structural roles of coexisting contradictions, and the historical reordering of those contradictions as a process develops. These ideas will later be transferred cautiously to scientific inquiry, where problems, methods, instruments, institutions, and criteria of evaluation can change together. The method here is therefore reconstructive rather than exegetical. Hegelian dialectics and historical materialism supply conceptual resources; the account of historically self-evolutionary AI developed later in the paper remains a distinct philosophical proposal.

DRAFT

2.1 Dialectical Development and Immanent Contradiction

A useful starting point is the difference between external opposition and immanent contradiction. An external opposition can be manufactured by placing two propositions, strategies, or viewpoints against one another. Contemporary language models are already capable of doing this fluently. Given a proposal, they can generate objections; given an argument, they can generate a counterargument; given two positions, they can produce a compromise. Such operations may improve deliberation, but they do not by themselves amount to dialectical development. The philosophical significance of Hegelian dialectics lies partly in the stronger claim that a determination can become unstable through limitations that arise from its own content and mode of organization. The transition is therefore not supposed to depend on an arbitrary contrary imported from outside. Hegel's dialectical method is commonly reconstructed as moving from a relatively fixed determination, through the disclosure of its one-sidedness or instability, toward a later determination generated through the development of the earlier one [17].

This point matters for artificial problem solving because a system that merely generates opposites can mistake rhetorical diversity for developmental necessity. Suppose a scientific model assumes that a phenomenon can be represented adequately by a particular set of variables. An artificial agent can easily produce an alternative model using different variables. A dialectical analysis asks a different question: does the original model encounter a limitation that follows from what it is trying to do? Perhaps improving predictive accuracy requires adding interactions that make the model unidentifiable with available observations. Perhaps increasing representational detail destroys the comparability that made the original model scientifically useful. Perhaps a method designed to eliminate one source of bias amplifies another because of the way the evidence is generated. In such cases, the relevant tension is not simply disagreement between two independently available opinions. It emerges from the development of the existing practice itself.

The distinction can be stated without assuming that every scientific difficulty is literally a contradiction in a strict logical sense. Hegelian scholarship itself distinguishes among logical contradiction, opposition, limitation, and forms of instability that are called dialectical in a broader sense [17]. For the purposes of this paper, the term *contradiction* will therefore be used cautiously to denote a structured incompatibility, tension, or mutually constraining relation whose persistence affects how a problem situation can develop. This usage is intentionally broader than the simultaneous truth of a proposition and its negation. It is also narrower than the everyday habit of calling any disagreement a contradiction. The relevant feature is developmental: the relation must matter to the reproduction, limitation, or transformation of the system under examination.

A historically self-evolutionary AI would consequently need more than adversarial generation. It would have to reconstruct why a problem state produces certain tensions and why some proposed alternatives genuinely respond to those tensions while others merely sit beside them. This requires attention to dependency. A contradiction is immanent, in the relevant sense, when it can be traced to the commitments, capacities, or relations that constitute the present state of inquiry. If a research programme simultaneously requires finer measurement and relies on an instrument whose operating principle imposes a resolution ceiling, the tension is generated by the programme's own conditions. If a field demands rapid replication while depending on extremely costly experimental infrastructure, the conflict is likewise embedded in the organization of inquiry. An artificial system that can identify such relations is doing something different from producing a list of pros and cons. It is attempting to locate the internal limits through which a present form of problem solving generates pressure for transformation.

This idea also changes the temporal unit of analysis. A static description records what positions or constraints exist at a given moment. Dialectical development asks how one state gives rise to another. The significant object is therefore not only a set of propositions but a process in which commitments are enacted, consequences accumulate, limitations become visible, and responses alter the conditions under which the original problem was posed. The later state cannot be understood fully by comparing two snapshots if the transition itself produced new evidence, concepts, institutions, or practices. Historical self-evolution accordingly requires process memory: not merely retention of earlier states, but representation of how the present state was generated from them.

DRAFT

2.2 The Transformative Logic of Sublation

The Hegelian concept of *Aufhebung*, conventionally translated as sublation, is important because it prevents dialectical development from being reduced either to rejection or to compromise. Hegel emphasizes the double sense in which a determination can be cancelled and preserved at once: a later form supersedes an earlier one while retaining elements that remain necessary within the new organization [17]. This structure is especially useful for thinking about scientific development because successful transformation often does not eliminate an earlier practice absolutely. A theory can be restricted to a domain in which it remains effective; an obsolete measurement technique can survive as a calibration standard; a conceptual distinction can be retained while the framework that originally housed it is abandoned. Historical development frequently reorganizes the function and scope of earlier achievements rather than simply deleting them.

For artificial intelligence, this creates a stricter requirement than generating a synthesis between two positions. A superficial system can take position A, produce position B, and write a balanced paragraph containing parts of both. Sublation demands a reasoned account of what must be relinquished, what must be preserved, and why the resulting organization addresses the limitation that generated the transition. Preservation is therefore selective and relational. What survives does so because it continues to perform a role within the transformed problem state. What is negated is not necessarily false in every respect; it may instead be inadequate as an organizing principle at the new stage.

Scientific examples illustrate the point. A model can remain locally useful after a broader theory changes the interpretation of its variables. A dataset gathered under an outdated classification scheme can remain valuable if its provenance permits recoding. A failed experiment can retain epistemic importance by delimiting a region of the hypothesis space. A research institution created for one technological era can be reorganized around a new infrastructure rather than dissolved. In each case, the historical successor contains traces of the predecessor, but those traces occupy a changed relation to the whole. The significance of an earlier result is therefore partly determined retrospectively by the developmental path that incorporates it.

This temporal structure matters for self-evolution because an AI that rewrites itself aggressively can lose precisely the resources that later become valuable. Recursive self-improvement framed purely as performance maximization can favor replacement: discard the worse strategy, retain the better one. Historical problem solving requires richer conservation. It needs mechanisms, whether computational or institutional, through which superseded approaches remain reconstructable, negative results remain interpretable, and abandoned branches can be revisited when later developments alter their relevance. The philosophical lesson of sublation is not that every earlier form deserves preservation. It is that development cannot be understood as a sequence of independent winners unless the relations of inheritance, restriction, and reinterpretation are also represented.

Sublation also introduces a distinction between solving a contradiction and transforming it. Some tensions disappear when one side is shown to be mistaken or when a technical obstacle is removed. Others are reorganized into a different problem. A new instrument can remove a measurement bottleneck while producing a data-interpretation bottleneck. A shared platform can solve a coordination problem while creating a governance problem. Automation can reduce labor scarcity while raising new questions about validation, provenance, or accountability. From a dialectical perspective, the success of an intervention can therefore be developmental precisely because it generates a new configuration of limits. A self-evolutionary AI should not interpret the reappearance of contradiction as evidence that progress has failed. The relevant question is whether the structure of the problem has changed in a way that enlarges, redirects, or clarifies what can be done next.

2.3 Historical Materialism and Situated Development

Dialectical development alone does not determine which transformations are materially possible. Historical materialism shifts attention from the movement of concepts considered in abstraction to the conditions under which forms of social life are produced and reproduced. In the 1859 *Preface to A Contribution*

DRAFT

to the Critique of Political Economy, Marx links historical transformation to the relation between developing productive forces and existing relations of production, arguing that social forms can become constraints on capacities that have developed within them [16]. The same text also contains a formulation especially relevant to the present argument: historical problems arise under conditions in which the material conditions of their solution are present or in formation [16]. Whatever one makes of the broader historical claims surrounding that formulation, it provides a powerful philosophical provocation for scientific problem solving: the existence of a question as a sentence does not imply the existence of the conditions required to investigate or establish an answer.

The present paper adopts this provocation without transferring Marx's social theory mechanically into the philosophy of science. Scientific inquiry is not simply a mode of economic production under another name, and laboratories, universities, datasets, or research communities should not be mapped one-to-one onto categories developed for the analysis of political economy. The transfer is structural and limited. Research also depends on capacities and relations: instruments must exist, skills must be distributed, evidence must be accessible, labor must be coordinated, infrastructures must be maintained, and institutions must authorize or obstruct particular forms of inquiry. Philosophy of science has long recognized that scientific knowledge is produced within social settings and that divisions of cognitive labor, trust, authority, funding, and institutional organization can have epistemic consequences [14]. A materialist orientation therefore asks how these conditions enter into the space of possible inquiry rather than treating them as external implementation details.

This shift is crucial for artificial intelligence because language models encounter scientific problems primarily through representations. Text makes it easy to flatten the difference between a problem that can be described and a problem that can be investigated. A model can discuss a hypothetical experiment whose required instrument does not exist, integrate datasets that cannot legally or technically be combined, propose validation that no available institution can perform, or assume observational access that would require decades of data collection. None of these proposals is necessarily conceptually incoherent. Their limitation is historical and material. The gap lies between the representable and the presently realizable.

A historically materialist orientation therefore changes the question posed to an artificial scientific agent. Instead of asking only what answer follows from the available literature, the system must ask what capacities make the answer testable, what relations govern access to those capacities, and which of those conditions can themselves be transformed. Some obstacles are conceptual; others are infrastructural. Some require a better theory; others require a sensor, a longitudinal dataset, a shared standard, a legal agreement, a new form of coordination, or simply time. Treating all of them as deficiencies in reasoning would produce category errors. Long-horizon problem solving depends on diagnosing the type of constraint because different constraints imply different developmental interventions.

Historical materialism also directs attention to reproduction. A condition matters not only because it exists at one moment but because institutions and practices continually recreate it. A fragmented data environment, for example, may persist because technical formats are incompatible, because organizations have divergent incentives, because privacy rules restrict transfer, because funding rewards local rather than shared infrastructure, or because authority over classification is contested. An intervention aimed at one surface manifestation can fail if the relations reproducing the fragmentation remain intact. This is another reason why historical self-evolution cannot be reduced to candidate search. The system must reason about the processes that regenerate the problem state.

2.4 The Material Organization of Knowledge Production

The distinction between productive forces and relations of production provides a further conceptual resource, but only if the analogy is handled carefully. In Marx's political economy, productive forces include productive capacities such as labor power, skill, tools, and means of production, while relations of production concern the social relations through which production is organized [16, 11]. For scientific inquiry, an analogous distinction can illuminate the difference between the capacities available for producing knowledge and the relations through which those capacities can actually be combined.

DRAFT

The productive capacities of inquiry can include trained researchers, experimental skill, theoretical techniques, computational resources, datasets, instruments, laboratories, software, standards, and accumulated bodies of knowledge. The relations of knowledge production include forms of access, ownership, authority, division of labor, collaboration, funding, publication, evaluation, governance, and institutional responsibility. This distinction is analytical rather than ontological. A dataset can function simultaneously as an epistemic resource, an institutional asset, and an object governed by legal and ethical relations. A scientific instrument embodies both technical capacities and organizational dependencies. The purpose of the distinction is therefore not to sort every element into a single box, but to ask whether a limitation lies primarily in what a research system can do or in how its capacities are organized and made mutually available.

This distinction helps explain why scientific capacity can increase while solvability remains constrained. A field may possess sophisticated instruments and extensive data but lack standards that permit comparison across sites. Multiple laboratories may have the technical capacity to reproduce an experiment but lack incentives or funding structures for replication. Researchers may possess complementary expertise but be separated by disciplinary institutions that make sustained collaboration costly. A highly capable AI may have access to the published record while lacking access to negative results, tacit experimental knowledge, proprietary measurements, or institutional decisions that shaped what entered the literature. In such cases, additional cognitive power does not automatically remove the relevant obstacle.

The reverse configuration is also possible. Relations can be favorable while productive capacities remain inadequate. A well-coordinated international collaboration can share all available data and still be unable to answer a question because measurement precision is insufficient. Open science infrastructure can remove access barriers without creating the longitudinal observations a causal claim requires. A community can agree on standards while lacking a theory that gives those measurements an interpretable relation to the phenomenon of interest. Historical diagnosis therefore needs to distinguish among different combinations of capacity and organization rather than attributing every developmental blockage to a single source.

For AI development, the implication is that a research-state representation should eventually include both kinds of condition. This does not require reproducing Marxian categories literally in software. It requires recognizing that scientific intelligence operates in a world where reasoning capacities are embedded in infrastructures and relations. An AI that recommends a research programme without representing those conditions risks proposing transitions that are logically describable but historically unreachable. Conversely, an AI that can identify a missing coordination structure or a missing instrument may contribute to problem solving even when it cannot yet provide the final scientific answer. The relevant unit of intelligence becomes the transformation of a problem-producing system, not only the generation of propositions within it.

2.5 Structural Asymmetry among Contradictions

Complex research situations ordinarily contain many tensions at once. A field can face limited data, expensive experiments, weak theory, incompatible standards, scarce expertise, publication pressure, replication difficulties, and institutional fragmentation simultaneously. Treating every tension as equally important produces an unstructured problem list rather than a developmental analysis. Historical reasoning requires attention to structural asymmetry: some contradictions can organize, reproduce, or condition several others, while some are secondary manifestations of a deeper bottleneck.

The present paper will use the language of *structurally central* and *derivative* contradictions for this distinction. The terminology is intentionally analytical rather than doctrinal. Structural centrality is not equivalent to moral importance, rhetorical salience, scale of harm, or subjective urgency. A contradiction is structurally central when changes in that relation are likely to reorganize several other constraints or alter the reachable paths of development. A derivative contradiction can still be severe, ethically important, or practically urgent. The distinction concerns developmental dependence rather than worth.

Consider a research programme in which theoretical models have become more complex while available observations remain sparse. The immediate symptoms may include unstable parameter estimates,

DRAFT

disagreements among teams, replication failures, and proliferation of incompatible assumptions. An AI could treat each symptom as an independent problem. A hierarchy-sensitive analysis would ask whether several of them depend on the same relation between model complexity and evidential resolution. If so, an intervention that improves measurement or reduces model indeterminacy may transform several apparent problems together. By contrast, optimizing publication workflows would leave the central epistemic relation intact even if it alleviated a visible secondary difficulty.

Structural centrality must be inferred rather than declared. Dependence can be causal, institutional, methodological, or historical. A contradiction may be central because it constrains many downstream decisions, because institutions continually reproduce it, because alternative research paths converge on the same bottleneck, or because removing it changes the set of questions that can be investigated. Different forms of centrality can conflict. A highly connected institutional bottleneck may be easier to change than a less connected physical limit; an ethically urgent constraint may not be the one whose transformation produces the largest expansion of research capacity. A historically self-evolutionary AI would therefore need to expose the basis of its hierarchy rather than simply output a ranked list.

This requirement is especially important because AI systems are vulnerable to salience substitution. Problems that are well represented in text, frequently discussed, or easily verbalized can appear more important than poorly documented material constraints. Historical-material analysis pushes in the opposite direction: it asks what relations actually condition development, including those that leave weak textual traces. The epistemic difficulty is substantial. Some of the most consequential constraints on inquiry are tacit, infrastructural, proprietary, or institutionally normalized. Their identification may require evidence that is unavailable to a language model. The framework developed here should therefore be understood as generating research obligations for AI systems, not as presuming that current models already possess sufficient access to satisfy them.

2.6 Temporal Reordering of Contradiction Structures

A contradiction hierarchy is historical because its ordering can change. Once a structurally central limitation is transformed, previously secondary tensions can become decisive, and successful interventions can generate new constraints that did not exist in the earlier state. This temporal reordering distinguishes developmental analysis from static bottleneck identification.

Scientific history offers many patterns of this kind even without imposing a single dialectical schema on them. Improvements in observation can shift the central difficulty from detection to interpretation. Increased computing capacity can shift a problem from numerical feasibility to data quality, model specification, or energy cost. Standardization can solve comparability problems while making disagreements over classification more consequential. Automation can increase experimental throughput until validation, provenance, and anomaly detection become the limiting conditions. In each case, the relation that organizes development changes because the intervention has changed the system in which later research occurs.

This possibility means that an artificial agent should not treat an identified contradiction as a permanent property of a field. The diagnosis must be indexed to a historical state and accompanied by an account of the conditions under which it may cease to be central. After intervention, the agent must reconstruct the dependency structure rather than assume that the previous hierarchy persists. The same applies to the opposing aspects within a contradiction. A research field can move from excessive theoretical freedom under scarce evidence to excessive evidential abundance under weak interpretive structure. The labels applied at the earlier stage do not determine the later developmental problem.

The temporal reordering of contradictions also limits the usefulness of optimization toward a fixed target. If an intervention changes the state sufficiently, the objective that justified the intervention may no longer be the objective that should govern the next stage. A system designed to maximize data collection can become maladaptive after data abundance has been achieved. A programme designed to increase model complexity can become counterproductive once interpretability or validation becomes the central constraint. Historical intelligence therefore requires sensitivity to stopping conditions and objective revision. Persisting with yesterday's solution after the contradiction structure has changed is one form of

DRAFT

developmental failure.

The same historical logic applies to epistemic methods and evaluative standards. The present argument does not attribute to historical materialism a ready-made theory of scientific methodology or objective functions. It draws a more limited implication: practices of inquiry are situated within material and social forms of scientific production, and the adequacy of a method or standard can depend on those forms. A procedure that organizes productive inquiry under one configuration can become restrictive after the capacities, scale, evidence, or organization of research have changed. In this sense, an inherited method or evaluator can move from enabling development to constraining further development without thereby having been simply erroneous at the earlier stage.

That possibility places method and evaluation inside the historical process rather than outside it. Revision still requires justification and cannot be triggered merely by poor performance. The relevant philosophical question is one of historical adequacy: whether the relation among method, criterion, object, and research state still supports the epistemic role for which the method or criterion was adopted. Later sections develop this requirement through methodological and evaluative lineage, so that transformation preserves the valid domain of earlier practices while making their historical limitations inspectable.

2.7 Branching Historical Development

A final foundation is required to prevent historical self-evolution from becoming a teleological optimization narrative. Dialectical and historical language is easily misread as implying that each stage necessarily produces a higher stage and that development converges toward a predetermined end. Such a picture would be unsuitable for scientific inquiry and dangerous as a design principle for autonomous AI. Histories of science contain contingency, branching, local regressions, institutional lock-in, abandoned programmes, rediscoveries, and parallel traditions. Philosophical work on scientific revolutions likewise emphasizes continuing disagreement about whether scientific development follows any single general pattern and about how discontinuity, rationality, and progress should be understood [18, 3]. The social organization of inquiry further means that different communities can sustain different trajectories because their evidence, instruments, standards, and institutional relations differ [14].

Historical self-evolution should therefore be represented as a branching process in which multiple transitions may be possible from a given state. An intervention can open some paths while closing others. A standard can create cumulative coordination while locking a field into a classification scheme that later becomes difficult to revise. A large infrastructure can enable unprecedented observation while concentrating resources around the questions that infrastructure can answer. An automated research platform can accelerate one style of experimentation while making alternative methodological traditions less competitive. These transformations can be productive and constraining at the same time.

The preservation of alternative paths consequently has epistemic value. A system that retains only the historically victorious branch loses information about unrealized possibilities and about the conditions under which a rejected approach might become relevant again. Negative results, discontinued models, minority methodologies, and failed experimental designs can function as resources for counterfactual analysis. Their value is not romantic attachment to discarded ideas. It lies in making the history of inquiry reconstructable enough to ask whether present constraints are necessary or path-dependent.

This branching perspective also places a limit on what the term *self-evolutionary* should imply. Evolution does not guarantee improvement, and historical transformation does not supply its own normative criterion. A system can become more capable in one dimension while becoming less corrigible, less plural, more resource-intensive, or more dependent on a narrow infrastructure. A historically self-evolutionary AI therefore requires evaluative questions that cannot be answered by developmental dynamics alone. Which paths should be preserved, which constraints should be transformed, and which forms of increased capability count as scientific progress remain normative and epistemological problems. Historical materialism and dialectics can help describe how a problem situation changes; they do not by themselves determine what changes ought to be pursued.

The philosophical foundations developed in this section thus support a specific conception of long-horizon artificial intelligence. Scientific problem solving is treated as historically situated transforma-

DRAFT

tion rather than timeless search. Contradictions matter when they arise from and structure a developing problem situation; transformation can preserve earlier achievements while altering their role; materially organized capacities and relations condition what can be investigated; contradictions differ in structural centrality and can exchange roles over time; and historical development branches rather than moving automatically toward a final state. The next section turns from these foundations to contemporary evolutionary AI and autonomous scientific research. Its purpose is to determine how much of this developmental picture is already represented in existing systems and where the remaining conceptual gap lies.

3 Evolutionary AI and Autonomous Scientific Research

The philosophical argument developed in the preceding section enters an AI literature in which several apparently relevant capabilities already exist. Contemporary systems can preserve populations of candidate programs, generate stepping stones, rewrite parts of their own code, debate alternative hypotheses, execute experiments, accumulate records of earlier attempts, and revise subsequent actions in response to empirical outcomes. It would therefore be misleading to contrast a historically self-evolutionary AI with a static image of present-day artificial intelligence. The relevant question is narrower and more demanding: which dimensions of historical development are already represented by evolutionary and autonomous-science systems, and which dimensions remain external to the computational process they optimize?

The literature reviewed here is organized around that question. The first group of systems uses evolutionary or search-based procedures to improve programs under explicit evaluators. The second extends adaptation to the agent itself. A third line of work seeks open-endedness by preserving diverse lineages, generating new environments, or allowing objectives to change through co-evolution. A fourth uses debate, critique, and adversarial interaction to diversify reasoning. A fifth automates increasingly large portions of scientific inquiry, from hypothesis generation to manuscript production and, in some cases, iterative laboratory feedback. These developments substantially weaken any simple claim that AI lacks evolution, memory, experimentation, or changing environments. At the same time, they make the remaining distinction more precise. Most systems still operate within a research regime whose basic problem ontology, evaluative interface, material infrastructure, and forms of access are supplied in advance or treated as environmental givens. Their evolutionary power is considerable, but the historical conditions under which a problem becomes formulable, researchable, and establishable are rarely themselves reconstructed as the object of development.

3.1 Evolutionary Optimization and LLM-Guided Program Search

A useful starting point is the recent convergence of large language models with evolutionary program search. FunSearch demonstrated that a pretrained language model can serve as a source of program variation while an external evaluator provides selection pressure. The system maintains a population of programs, preferentially samples high-performing candidates, prompts the language model to produce variations, evaluates the resulting code, and stores successful programs for later use [19]. This arrangement enabled new results in extremal combinatorics and improved heuristics for online bin packing. The significance of FunSearch is not only its performance. It establishes a general pattern in which a generative model does not need to be intrinsically reliable if candidate outputs can be subjected to a sufficiently discriminating external evaluation procedure.

AlphaEvolve expands this pattern from the evolution of a limited program component toward the evolution of more substantial algorithmic code. It combines language-model generation, automated evaluators, and an evolutionary program database, retaining promising solutions as material for subsequent generations [7]. Its reported applications include algorithmic improvements in mathematics, computing infrastructure, chip design, and AI training. The underlying epistemic structure remains especially clear: a candidate is generated, executed, measured, and selected according to objective or quantifiable criteria. This structure is powerful precisely where candidate quality can be evaluated systematically. It

DRAFT

allows search to exceed the one-shot capabilities of a language model by embedding generation within a cumulative selection process.

The 2026 Empirical Research Assistance system, or ERA, makes the dependence on an evaluative interface still more explicit. ERA uses a language model and tree search to write and revise empirical software with the goal of maximizing a defined quality metric, and has been applied to scorable scientific tasks in bioinformatics, epidemiology, geospatial analysis, forecasting, and other domains [2]. The system can import research ideas from papers, textbooks, and search results, recombine successful implementations, and systematically explore candidate solutions at a scale difficult for human researchers. In this sense, it already performs a form of cumulative methodological development: earlier successful ideas are retained as resources, later candidates build on them, and empirical scores determine which branches deserve further investment.

These systems are directly relevant to the present paper because they show why "evolution" cannot be the proposed novelty. Variation, inheritance, selection, population diversity, and cumulative improvement are already productive components of AI-assisted discovery. They also demonstrate that an artificial system need not solve a difficult problem by deriving a complete answer in one inference. It can instead produce a succession of partial improvements whose cumulative effect surpasses the initial model.

Their limits, however, identify the first boundary of the historically materialist proposal. FunSearch begins with an evaluation function and usually with a user-specified program skeleton; AlphaEvolve relies on automated evaluators that can score proposed algorithms; ERA explicitly targets tasks for which a quality metric can be defined [19, 7, 2]. The central developmental uncertainty is therefore located primarily in the candidate solution, not in the constitution of the problem or the conditions under which candidate quality becomes measurable. The evaluator may be complex, but it functions as a relatively stable interface between the search process and the world. The historical-material problem developed in this paper begins when that interface is itself incomplete: when the field lacks an adequate measurement, when the relevant categories are still disputed, when access to evidence depends on changing institutions, when a necessary instrument does not yet exist, or when the criteria by which a result ought to count as progress are transformed by the inquiry itself.

This distinction should not be exaggerated into an opposition between "mere optimization" and "real science." Program-search systems can produce genuine scientific knowledge, and their evaluators can encode nontrivial empirical or mathematical structure. The narrower claim is that they usually receive a substantial part of the problem's epistemic organization in executable form. A historically self-evolutionary account asks what an AI should do when part of its task is to discover, construct, or reorganize the very conditions that would make such evaluation possible.

3.2 Recursive Self-Improvement and Self-Modifying Agents

A second line of research moves the evolving object from the candidate solution toward the problem-solving agent itself. The Gödel Agent framework treats agent logic and behavior as revisable rather than fixed, allowing an LLM-based system to modify its own implementation in pursuit of high-level objectives [25]. The Darwin Gödel Machine develops this idea through open-ended empirical self-improvement. Instead of requiring a formal proof that a self-modification is beneficial, it generates modified coding agents, evaluates them on software-engineering benchmarks, and preserves a growing archive of successful and diverse variants [26]. Because later agents can be generated from earlier members of the archive rather than from a single incumbent, the process can exploit stepping stones that would be lost under strictly greedy replacement.

This work is especially important for the argument of this paper because it already contains several properties that a superficial account might otherwise claim as novel. The Darwin Gödel Machine has historical memory in the weak but significant sense of an archive; it has branching lineages rather than one linear sequence; its modifications can change the mechanisms through which later modifications are produced; and earlier variants can become useful as intermediate stages even if they are not globally dominant at every moment. Self-evolution is therefore not merely a metaphor for repeated prompting.

DRAFT

It can refer to an empirical process in which the architecture of agency is itself subject to cumulative variation and selection.

Yet the object of self-modification also clarifies a second distinction. Recursive self-improvement asks how an agent can become a better solver under some evaluative regime. Historical self-evolution, as proposed here, asks in addition how the agent represents the historical transformation of the situation being solved. An agent may rewrite its code, improve its tool use, extend its context management, and refine its internal review mechanisms while continuing to encounter the external problem through the same benchmark interface. In such a case, the solver evolves while the ontology of the task and the structure of its evaluation remain comparatively stable.

This is not a defect in systems such as the Darwin Gödel Machine. Their research objective is different, and fixed or well-controlled benchmarks are necessary for determining whether self-modification has actually improved performance. The distinction matters because scientific development often contains a second-order problem that cannot be represented by improved agent performance alone. A field can acquire more capable researchers and more powerful computational tools while remaining blocked by unavailable measurements, inaccessible populations, proprietary evidence, unsuitable institutional arrangements, or conceptual categories that make relevant phenomena difficult to observe. Increasing the capability of the problem solver does not automatically transform these relations.

The difference can be expressed as a difference in the evolving state. Recursive self-improvement primarily changes properties of the artificial agent. A historically materialist analysis also tracks changes in evidence, instrumentation, methods, institutional arrangements, access, division of scientific labor, and the dependency structure among unresolved problems. This broader state can change because of actions taken by the AI, but it is not reducible to the AI's code or policy. A laboratory instrument produced through a research programme, a shared dataset established through international coordination, or a new evidential standard accepted by a community may expand scientific reachability even if the underlying agent model remains unchanged.

The distinction also affects the interpretation of failure. In recursive self-improvement, failure can become training material for producing a better agent. In historically developing science, failure may instead reveal that the problem has been mis-specified, that the measurement process cannot discriminate among hypotheses, or that the current social organization of inquiry is reproducing the bottleneck. In these cases, "improve the solver" and "transform the problem situation" are different responses. A future autonomous scientist may need both.

3.3 Open-Ended Evolution and Evolutionary Stepping Stones

Open-ended learning provides a closer precedent for the idea that problems themselves can evolve. The Paired Open-Ended Trailblazer, or POET, was explicitly designed to generate environmental challenges at the same time as it generates agents capable of solving them. Rather than optimizing one agent against one fixed task, POET maintains a collection of environment-agent pairs, creates new environments, transfers candidate solutions across them, and uses apparently indirect stepping stones to reach behaviors that direct optimization does not readily discover [20]. This is philosophically important because it breaks the assumption that the sequence of problems must be fully specified before learning begins. The curriculum is partly generated by the system itself.

The open-endedness literature therefore anticipates one of the central intuitions of historical self-evolution: an intermediate state can be valuable because of what it makes reachable later. A solution that is not the shortest path toward a final target can become indispensable if it opens another branch of the search space. The Darwin Gödel Machine imports this stepping-stone logic into recursive agent improvement by maintaining an archive of diverse agents rather than collapsing the population into a single winner [26]. The AI Scientist similarly uses archives of research ideas, and its later template-free form incorporates tree search to preserve multiple experimental branches before selecting promising continuations [15].

Digital Red Queen pushes the comparison further by making the effective objective itself dynamic. In its Core War environment, each newly evolved program must compete against predecessors, so the

challenge faced by later agents changes as the population changes [10]. The work is explicitly motivated by the observation that biological, technological, and cultural evolution often occurs under continually shifting pressures rather than on a fixed fitness landscape. It therefore supplies a useful counterexample to any claim that contemporary evolutionary AI necessarily assumes a static environment.

These systems narrow the conceptual gap considerably. Problems can be generated, lineages can branch, objectives can change through interaction, and solutions can act as stepping stones toward capabilities that were not directly optimized at the beginning. The historically materialist proposal must therefore add something more specific than openness, path dependence, or environmental change.

The difference lies in what produces the changing environment and how that change is interpreted. In POET, the generator and admission criteria define a formal space of possible environmental mutations. In Digital Red Queen, the changing objective is produced by adversarial co-evolution among programs in a bounded simulation. These are genuinely dynamic problem spaces, but the mechanisms through which new problems become possible remain part of the designed computational regime. The system does not ordinarily need to reconstruct how laboratories were funded, why particular data are absent, how an instrument's material limitations constrain theory, why one classification became institutionally dominant, or how a publication norm alters the evidential record available to future inquiry.

Historical-material development is therefore not synonymous with a nonstationary fitness landscape. A changing objective can remain ahistorical in the stronger sense used in this paper if the process does not represent the causal and institutional history through which the objective became salient. Conversely, a relatively stable scientific objective can contain substantial historical development if the available means of pursuing it, the evidence considered admissible, or the relations organizing inquiry change over time. The relevant question is not only whether the environment changes but whether the AI can reconstruct the generative relations that make those changes intelligible and actionable.

This point also explains why a simple import of open-ended evolution into scientific AI would be insufficient. Scientific research cannot safely treat every newly generated problem as equally worth pursuing. Resource constraints, ethical limits, human values, public obligations, and epistemic dependencies shape which branches should be explored. Open-endedness supplies mechanisms for preserving possibility; it does not by itself supply an account of the material and normative conditions under which one research trajectory should be developed rather than another. The framework proposed later in this paper therefore treats plural branching as necessary but not self-justifying.

3.4 Multi-Agent Debate and Adversarial Reasoning

Another adjacent literature uses interaction among language-model agents to improve reasoning. Multi-agent debate systems allow multiple instances to propose answers, inspect one another's reasoning, criticize alternatives, and revise their conclusions over several rounds. Early work reported gains in mathematical reasoning, strategic reasoning, and factuality from such iterative debate [5]. Later research has examined the importance of diversity among agents. Diverse Multi-Agent Debate, for example, explicitly seeks different reasoning approaches rather than merely assigning different personas to otherwise homogeneous agents, with the aim of escaping shared "mental sets" [13].

For the present paper, debate systems matter because they demonstrate that contradiction-like interaction can be computationally useful. An artificial reasoner need not maintain one uninterrupted chain of thought. It can externalize alternatives, expose disagreements, and use criticism to revise a candidate conclusion. AI Co-Scientist builds related mechanisms directly into scientific hypothesis generation: specialized agents generate, reflect on, rank, debate, and evolve hypotheses within an asynchronous multi-agent framework [9]. The resulting process is much closer to scientific argument than a single model producing a final proposal without opposition.

There is also emerging work that explicitly adopts dialectical vocabulary. Recent proposals have evaluated language-model reasoning through the interaction of thesis, antithesis, and synthesis, arguing that static answer correctness does not fully capture whether a model can integrate conflicting ideas into a more developed conclusion [1]. Such work is directly relevant to the philosophical motivation of this

DRAFT

paper, although it uses dialectics primarily as a reasoning or evaluation structure rather than as an account of historical-material development.

The crucial distinction is between adversarial diversity and immanent contradiction. Debate methods generally obtain diversity by generating alternative answers, assigning different strategies, or arranging competition among agents. This can expose errors and reduce premature consensus. A dialectical analysis in the stronger sense developed in Section 2 asks why an opposition arises from the commitments and conditions of the problem state itself. The opposing position is not valuable merely because it differs. Its significance depends on whether it expresses a limitation generated by the current organization of inquiry.

This creates at least three unresolved requirements. First, an AI needs to distinguish a genuine developmental tension from an arbitrarily generated counterposition. A model can always invent a contrary argument, but rhetorical symmetry does not show that the problem has a dialectical structure. Second, it must determine whether several visible disagreements are independent or whether they are symptoms of a more structurally central contradiction. Multi-agent debate ordinarily evaluates arguments at the level at which they are presented; it does not necessarily reconstruct a hierarchy of material dependencies behind them. Third, a dialectical transformation should alter the problem representation when the contradiction is mediated. Debate systems usually seek a better answer to the original question, whereas historical self-evolution may require concluding that the original question is no longer the right unit of analysis.

The literature also cautions against assuming that interaction itself guarantees higher-order reasoning. Controlled studies of multi-agent debate have found that outcomes depend strongly on the intrinsic reasoning ability and diversity of the participating agents, while majority pressure can suppress correction and apparent deliberation can collapse into forms of ensembling or voting [22]. These findings reinforce the philosophical caution adopted here. A future historically self-evolutionary AI cannot obtain dialectical intelligence merely by placing several language models into argumentative roles. The system needs criteria for relating disagreements to the causal and historical structure of the problem situation.

3.5 Autonomous Hypothesis Generation and Scientific Agents

The most direct neighboring literature is autonomous science. Systems in this area increasingly automate not one scientific task but a sequence extending across literature review, hypothesis generation, experimental design, execution, analysis, and communication. Surveys now describe this shift as a movement from AI tools for science toward "agentic science," in which artificial systems participate in iterative research workflows with increasing autonomy [21, 23].

AI Co-Scientist represents one important model. It is conditioned on scientist-provided research objectives and previous evidence, then uses specialized agents and a tournament-style evolutionary process to generate, criticize, rank, and refine hypotheses [9]. The system has been evaluated in biomedical applications including drug repurposing, target discovery, and hypotheses about bacterial evolution. Its importance for the present argument lies in the combination of scientific grounding and explicit hypothesis evolution. It does not simply brainstorm alternatives; it allocates computational effort across candidate hypotheses and iteratively improves them through internal critique and ranking.

The AI Scientist extends automation across a larger research lifecycle. Its current form can generate research directions, search literature, write and execute experimental code, analyze results, produce figures, draft a manuscript, and subject the manuscript to automated review. A 2026 Nature report describes both scaffolded and more open-ended modes, including agentic tree search for experimental exploration and an archive of research ideas [15]. The system therefore already incorporates several features that resemble long-horizon research: memory, branching experiment trajectories, stopping conditions, automated debugging, negative results, literature-based novelty checks, manuscript production, and evaluation.

The significance of this work is easy to underestimate if it is described merely as "paper generation." Its scientific workflow changes the location of human intervention. In conventional AI-assisted research, researchers formulate the project and delegate isolated tasks. In autonomous-science systems,

DRAFT

the artificial agent can control substantial portions of the transition from one research state to another. The generated experiment changes the evidence available to the next step; the observed result changes the planned analysis; the analysis informs what is written and, in more iterative systems, what is investigated next.

Recent systems push this feedback structure beyond computational machine-learning experiments. Coscientist demonstrated an LLM-based agent capable of planning and orchestrating chemistry experiments through access to laboratory automation, documentation, web search, and code execution [4]. Robin, reported in 2026, connects literature-based hypothesis generation with autonomous analysis of biological experimental data in an iterative lab-in-the-loop cycle. It generates therapeutic hypotheses, proposes experiments, interprets returned laboratory results, and uses those results to generate updated hypotheses [6]. The system therefore contains a genuine empirical feedback loop in which later questions depend on evidence produced by earlier interventions.

These developments directly challenge any strong assertion that contemporary AI only operates synchronically. Scientific agents can maintain experimental journals, revise hypotheses after failure, preserve prior branches, and use newly produced evidence to determine subsequent action. AutoResearchClaw goes further in explicitly carrying lessons across research runs, combining multi-agent debate, failure-driven pivoting, verifiable result reporting, human intervention modes, and cross-run evolution intended to transform previous mistakes into future safeguards [12]. Taken together, the literature shows a rapid movement toward systems whose behavior unfolds through extended research histories rather than isolated prompts.

The remaining question is therefore not whether AI can participate in temporal scientific processes. It can. The question is what kind of history those processes represent.

Most autonomous-science systems still begin from a substantial externally specified research frame. AI Co-Scientist is aligned to scientist-provided research objectives. The AI Scientist operates within broad research domains and implements a recognizable research lifecycle with explicit stages and evaluative procedures. Coscientist operates through an available laboratory interface whose instruments and APIs already define what actions can be taken. Robin can revise hypotheses in response to biological experiments, but the institutional and material organization of the experimental loop is supplied as infrastructure rather than treated as an object of historical transformation. ERA targets tasks whose success is expressible by a quality metric. These systems may adapt brilliantly within their research regimes without yet representing the genesis and possible transformation of those regimes themselves.

This limitation is not absolute. In practice, autonomous agents can decide to collect different data, build new software, propose new experiments, or search for new literature, and those actions do modify the conditions of later inquiry. The distinction is instead one of explicit representation. A historically materialist system would need to treat conditions such as instrumentation, data access, institutional coordination, labor allocation, evidential standards, and conceptual classification as first-class parts of the problem state. It would ask whether changing one of those conditions is itself the appropriate scientific intervention. Current autonomous-science architectures usually emphasize movement through the scientific workflow; the present proposal emphasizes transformation of the historically constituted field in which that workflow operates.

3.6 Closed-Loop Experimentation and the Expansion of Research Reachability

Closed-loop experimentation deserves separate attention because it comes closest to changing what is knowable through action. An agent that only manipulates text can propose hypotheses but cannot create new empirical evidence. An agent connected to software execution, simulation, robotic laboratories, or human experimental collaborators can intervene in the world and thereby alter the evidence available to future reasoning. This transition is essential to any historically materialist account because the material conditions of inquiry cannot be reduced to stored propositions.

Coscientist provides an early demonstration of this transition by connecting language-model planning to laboratory automation in chemistry [4]. Robin provides a stronger example of iterative epistemic transformation: one round of hypothesis generation motivates experiments, the resulting data are analyzed,

and the analysis produces a new round of hypotheses and follow-up experiments [6]. The importance of such systems is not that they perfectly reproduce human scientific practice. It is that they instantiate a loop in which actions generate evidence that did not exist at the beginning of the process.

From the perspective of historical self-evolution, this is the beginning of research-state transformation. If an intervention produces a new dataset, validates a measurement procedure, establishes a failed region of the search space, or reveals a previously unnoticed variable, the next scientific state is not merely a better internal belief state of the agent. The publicly or institutionally available evidence has changed. In principle, autonomous systems could extend this process to constructing reusable tools, standardizing data, designing measurement devices, or coordinating distributed experiments. Such actions would increase the future reachability of questions beyond the immediate hypothesis that motivated them.

The literature nevertheless reveals a persistent asymmetry between experimental action and historical reconstruction. Closed-loop systems are typically given the laboratory, simulation, benchmark, or software environment through which action becomes possible. They reason over the opportunities made available by those infrastructures but do not normally ask how the infrastructures themselves came to define the field's possibilities. Nor do they generally compare scientific interventions according to how they transform the long-run space of researchable questions. A successful experiment is valued primarily because it tests or improves a hypothesis, not because it changes the material organization of future inquiry.

This distinction suggests why "starting from an easy problem" has a stronger meaning in the present paper than ordinary curriculum learning or experiment sequencing. An intermediate research problem may be valuable because solving it constructs an instrument, dataset, conceptual distinction, protocol, or coordination mechanism that changes the feasible set of later investigations. The desired system therefore needs to reason over two levels simultaneously: the immediate epistemic value of a research action and its effect on the future production of knowledge. Current closed-loop science provides components for this possibility, but generative transformation of research reachability is not yet a standard objective of autonomous scientific agents.

3.7 Current Limits of Evolutionary Scientific AI

The literature reviewed above permits a more careful statement of the conceptual gap. None of the following can plausibly be claimed as unique to the historically self-evolutionary proposal: iterative refinement, evolutionary search, self-modification, archive-based memory, stepping stones, branching exploration, dynamically changing objectives, multi-agent criticism, hypothesis evolution, automated experimentation, learning from failure, or feedback from newly produced evidence. Contemporary AI already implements each of these in meaningful forms [20, 26, 10, 9, 15, 6].

The remaining gap concerns the level at which development is represented. Current evolutionary systems usually optimize candidates, agents, hypotheses, programs, or experiment paths. Even when the environment changes, the mechanisms that define admissible variation and evaluation are largely part of the designed system. Current scientific agents traverse increasingly complete research workflows, but the workflow's material and institutional preconditions usually appear as available tools, APIs, datasets, laboratories, literature services, benchmark metrics, or human-provided objectives. These are powerful resources, yet they are often treated as background rather than as historically produced relations that can themselves constrain scientific development.

A historically materialist analysis asks the AI to make those background conditions explicit. The agent should be capable, at least in principle, of recognizing that a difficult scientific problem may be blocked primarily by a missing instrument rather than a missing hypothesis; by fragmented access to data rather than insufficient inference; by an evaluative convention that rewards the wrong proxy; by a division of labor that prevents relevant expertise from interacting; by a category scheme that obscures the phenomenon; or by a research infrastructure whose design makes certain questions cheap and others practically invisible. These conditions are neither simply "context" nor merely additional features of a benchmark. They partly constitute what can be researched at the present historical stage.

DRAFT

This perspective also changes the role of evaluation. In FunSearch, AlphaEvolve, and ERA, evaluability is a source of power: reliable scoring allows enormous search. In The AI Scientist, automated review provides a scalable selection mechanism for generated research. In AI Co-Scientist, ranking and tournament evolution help allocate effort among hypotheses. These evaluators can themselves be revised by engineers, and some systems use multiple forms of evaluation. Yet the broader research process can encounter situations in which the appropriate criterion of progress is itself uncertain or historically unstable. A metric can become a target that distorts inquiry; an evidential standard can become obsolete after a new measurement technology; a benchmark can cease to represent the scientific purpose for which it was introduced. Historically self-evolutionary AI therefore requires not evaluator-free search but a principled capacity to diagnose when an evaluator has become part of the problem.

A similar issue arises with contradiction. Multi-agent debate and hypothesis criticism already expose competing positions. What remains underdeveloped is a model of contradiction hierarchy tied to the material dependency structure of inquiry. The system must ask not only which argument defeats another but which tension reproduces several others, which bottleneck limits the reachable research state, and whether resolving a visible dispute leaves the deeper condition unchanged. This is where the Hegelian concern with immanent limitation and the historical-material concern with conditions of development enter as more than rhetorical vocabulary.

Finally, present autonomous-science systems are generally oriented toward acceleration: producing better hypotheses, testing more candidates, writing better empirical software, conducting more experiments, or completing more of the research lifecycle with less human intervention. These are legitimate and potentially transformative aims. Historical self-evolution adds a different temporal criterion. Sometimes the appropriate next step is not the fastest route to an apparent answer but the construction of conditions under which a question can later be answered responsibly. A negative result, a standardized dataset, a measurement protocol, a conceptual clarification, or an institutional coordination mechanism can have low immediate performance value and high developmental value. An AI oriented toward the history of solvability would need to recognize this difference.

The literature therefore points not to an absence of evolutionary intelligence but to an unfinished enlargement of its object. AI research has moved from optimizing parameters to evolving programs, from evolving programs to modifying agents, from fixed tasks to generated environments, from isolated responses to debate, and from individual scientific subtasks to increasingly autonomous research cycles. The conceptual gap that remains is the status of the conditions under which such research becomes possible. The next section therefore begins the positive argument of this paper by examining scientific solvability as historically conditioned: evidence, methods, instruments, labor, institutions, access relations, and evaluative practices are treated not merely as background resources but as variables that can determine which problems are presently researchable and which remain only conceivable.

4 Historical Conditions of Scientific Solvability

The preceding sections established two points that now need to be brought together. The philosophical discussion treated development as transformation within historically situated conditions rather than as movement through an abstract sequence of better answers. The literature review then showed that contemporary evolutionary and autonomous-science systems already implement substantial forms of iteration, self-modification, open-ended search, debate, hypothesis evolution, and experimental feedback. The remaining question is therefore not whether artificial intelligence can evolve, but what it would mean for an artificial scientist to reason about the historical conditions under which a scientific problem becomes solvable. This section develops that question before introducing any detailed model of a research state. It proceeds from the historical situatedness of scientific problems, through the material conditions of researchability, to distinctions among conceptual possibility, researchability, and establishability. It then separates productive capacities of inquiry from the relations through which those capacities are organized, and concludes by treating solvability itself as a historical variable.

DRAFT

4.1 Scientific Problems in Historical Context

Scientific problems are often written as if they were temporally stable objects. A question can be stated in a form that appears independent of the moment at which it is asked: What is the structure of a molecule? What causes a disease? How does a physical system behave under specified conditions? Can a mathematical conjecture be proved? This linguistic stability can obscure a more consequential historical instability. The question that can be written at one time need not be the question that can be investigated, discriminated, or established at that time. A problem exists simultaneously as a conceptual formulation and as a position within a historically developed field of possible inquiry.

This distinction matters because scientific questions inherit dependencies from previous developments. A problem may presuppose a vocabulary capable of distinguishing relevant phenomena, a measurement procedure capable of making those distinctions observable, a body of prior results that narrows the plausible explanations, and an institutional setting capable of sustaining the required work. When one of these dependencies is absent, the problem can remain intelligible at the level of language while lacking a viable path toward evidential resolution. The historical character of science therefore enters before any answer is generated. It enters into the constitution of the problem itself.

The point is compatible with, but not reducible to, familiar observations about paradigms and scientific revolutions. Historical philosophy of science has long emphasized that problems, standards, and disciplinary significance can change together rather than developing against an invariant background [18, 3]. The present argument adopts a narrower implication. Even without committing to a general theory of scientific revolutions, an account of autonomous scientific intelligence must distinguish a problem's semantic availability from its historically available means of investigation. A model may be able to formulate an elegant research question by drawing on linguistic traces of many periods at once while the present research environment still lacks the evidence, instruments, coordination, or conceptual stabilization required to investigate it responsibly.

This is one reason why large language models create a distinctive epistemic situation. Their training data compress textual products from heterogeneous historical stages into a common inferential environment. A model can juxtapose a contemporary statistical technique, an older philosophical distinction, a newly developed instrument, and a concept imported from another discipline without reproducing the historical work through which those resources became reliable. Such recombination can be scientifically productive. It can also hide dependencies. The ease with which a model can state a mature-looking proposal does not show that the surrounding field possesses the conditions needed to test, interpret, or establish it.

The relevant historical unit is therefore not simply a date. To call a problem historically situated is to say that its scientific status depends on a configuration of prior results, available evidence, technical capacities, institutional arrangements, and interpretive practices. Two communities at the same chronological time can occupy different effective research situations because they possess different instruments, data access, expertise, infrastructures, or standards. Conversely, communities separated in time can confront structurally similar limitations. Historical situatedness is thus relational: it concerns the state from which inquiry proceeds, not merely the calendar position of the inquirer.

For self-evolutionary AI, this implies that problem representation cannot terminate at a proposition to be solved. The system must eventually reconstruct the dependencies that make the proposition scientifically actionable. A question may require a measurement capability that does not exist, a dataset that cannot yet be collected at sufficient quality, a theoretical distinction whose observable consequences remain unclear, or a coordination structure that would permit evidence to accumulate across sites. These are not secondary logistical details added after the intellectual work. They can determine whether the intellectual work has an evidential path at all.

4.2 Material Conditions of Researchability

A historical-material orientation makes these dependencies explicit by treating inquiry as an activity conducted through material and social arrangements. Scientific reasoning occurs through bodies, instru-

DRAFT

ments, laboratories, observatories, computers, archives, databases, software, standards, organizations, and forms of coordinated labor. Even highly abstract work depends on infrastructures of preservation, communication, verification, and education. Social epistemology similarly emphasizes that the production of scientific knowledge is distributed across communities and shaped by trust, authority, division of cognitive labor, criticism, and institutional organization [14]. A scientific AI that models only propositions and hypotheses therefore represents only part of the system through which knowledge becomes possible.

The most obvious material conditions are instruments and evidence. A phenomenon can become scientifically tractable when a measurement device makes previously inaccessible variation observable. A new detector can expand the domain of events that leave usable traces; a sequencing technique can change the granularity at which biological variation is studied; a computational platform can make a previously impractical simulation routine; a standardized dataset can make comparisons possible across research groups. In each case the scientific effect is not exhausted by an increase in efficiency. The intervention can alter what distinctions can be drawn and which hypotheses can be discriminated.

Data availability has a similar historical structure. The statement that a field "has data" is too coarse. Researchability depends on provenance, resolution, comparability, representativeness, accessibility, and the relation between recorded variables and the phenomenon of interest. Data that exist but cannot be legally accessed, technically integrated, or interpreted across incompatible standards may not function as available evidence for a given project. Conversely, the creation of a shared protocol can transform scattered observations into a cumulative evidential resource. For an autonomous system, access conditions and interoperability therefore belong to the epistemic description of the problem, not merely to project management.

Human and artificial labor also enter materially. Scientific work requires the allocation of attention, expertise, maintenance, interpretation, and responsibility. A research problem can be blocked because the relevant skills are absent from one institution, because disciplinary boundaries prevent collaboration, because the cost of maintaining an instrument exceeds available resources, or because routine labor required to produce high-quality evidence is institutionally undervalued. The increasing role of AI does not remove these dependencies. It changes their organization. Models require compute, tools, interfaces, permissions, supervision, and access to environments in which their outputs can be tested. A self-evolving scientific agent is therefore embedded in a productive arrangement even when much of its reasoning is automated.

Institutional conditions are equally consequential. Funding structures affect which long-horizon projects can be sustained. Publication practices influence whether negative results and replication work enter the public record. Ownership and licensing arrangements shape access to data, instruments, and software. Ethical and legal constraints appropriately delimit which interventions may be performed. Evaluation systems influence which contributions are rewarded and therefore which capacities are reproduced. None of these conditions mechanically determines scientific content, and the present paper does not claim that every epistemic disagreement reduces to institutional structure. The narrower claim is that such relations can alter the reachable paths through which a scientific problem may be investigated.

This is where a historically materialist orientation differs from a generic appeal to "context." Context can remain an indefinitely large collection of background facts. Material conditions of researchability are more specific: they are conditions whose presence, absence, organization, or transformation changes the feasible processes through which evidence and explanation can be produced. An AI capable of historical problem solving would need to identify which of these conditions are causally relevant to the present bottleneck. Listing all institutions, tools, or resources would be as unhelpful as listing all hypotheses. The analytical task is to reconstruct the relations that presently constrain or enable development.

The distinction also prevents an overly mentalistic conception of scientific progress. If a system cannot answer a question, the appropriate response need not be to improve its reasoning model. The limitation may lie outside the reasoner: insufficient measurement, inaccessible evidence, absent coordination, an underdeveloped standard, or a missing experimental platform. A self-improving agent that treats every failure as evidence of its own cognitive insufficiency risks optimizing the wrong object. His-

DRAFT

torical self-evolution requires the possibility that the solver correctly diagnoses itself as adequate for the current stage while recognizing that the surrounding research conditions must first change.

4.3 Conceptual Possibility, Researchability, and Establishability

The preceding discussion motivates a three-way distinction that is central to the rest of the paper: conceptual possibility, researchability, and establishability. These categories are related but should not be collapsed. A proposition or question is conceptually possible when it can be formulated coherently enough to enter reasoning. It is researchable when there exists a practically and methodologically available path for producing information relevant to its resolution. It is establishable when the available or producible evidence and inferential practices are sufficient to support a warranted scientific conclusion under the relevant standards of the field.

Conceptual possibility is the broadest category. Scientific imagination routinely outruns available evidence. Researchers can formulate mechanisms whose decisive observations are not yet possible, propose entities that current instruments cannot detect, or state mathematical and computational problems for which no tractable method is known. This surplus of conception over present researchability is not a defect. It is one source of scientific direction. The problem arises when an AI treats the ability to formulate a proposal as evidence that the proposal is already ready for scientific resolution.

Researchability adds a path condition. A problem becomes researchable when available methods and resources permit interventions, observations, derivations, simulations, or other procedures that can meaningfully change the evidential state. The path need not guarantee success. Researchability means that inquiry can proceed in a way capable of producing discriminating information. A difficult experiment whose outcome is uncertain can be researchable; a speculative theory with no current means of distinguishing its predictions from alternatives may remain conceptually meaningful while being weakly researchable.

Establishability is stronger again. A research path can exist while the resulting evidence remains insufficient for a stable conclusion. Measurements may be too noisy, samples too limited, models underidentified, or alternative explanations too difficult to discriminate. A field can therefore make genuine research progress without yet possessing the conditions of establishment. This temporal gap is ordinary in science. It becomes especially important for autonomous AI because fluent synthesis can make provisional structures appear more settled than the evidential state warrants.

The distinction can be written schematically as

$$\mathcal{E}_t \subseteq \mathcal{R}_t \subseteq \mathcal{C}_t,$$

where $\mathcal{C}_t$ denotes what is conceptually formulable at a given research state, $\mathcal{R}_t$ what is presently researchable, and $\mathcal{E}_t$ what is presently establishable. The notation is deliberately modest. These sets need not have sharp boundaries, and membership can be graded, contested, or community-dependent. Their purpose is to prevent a category error: the existence of a path in conceptual space does not imply the existence of a path in evidential practice.

Historical development alters these relations. A new instrument can move a question from $\mathcal{C}_t$ into $\mathcal{R}_{t+1}$. A standardized protocol can move a previously fragmented body of observations closer to $\mathcal{E}_{t+1}$. A conceptual distinction can have the same effect by making the relevant variables identifiable. Conversely, loss of infrastructure, restricted access, institutional collapse, or degradation of an archive can make previously researchable questions less accessible. Reachability is therefore neither monotonic nor guaranteed.

This framework also clarifies epistemic stage-skipping. Stage-skipping occurs when an AI generates a claim or theoretical organization that presupposes transitions not yet realized in the surrounding research state. The resulting proposal may be insightful and may even eventually prove correct. The problem is not necessarily logical error. It is the concealment of missing epistemic links. A historically aware system should be able to say that a proposed conclusion belongs to a possible later stage while identifying the intermediate work required before the conclusion can be responsibly established.

DRAFT

Such a diagnosis changes the function of planning. Instead of decomposing a hard problem only into smaller logical subtasks, the AI may need to identify state-changing prerequisites. The next task could be to develop a measurement method, construct a benchmark, standardize a vocabulary, collect a missing class of observations, test whether a distinction is empirically meaningful, or organize a collaboration capable of combining distributed expertise. These actions can look peripheral when judged against the final question, yet they can be constitutive of the path by which that question becomes scientifically answerable.

4.4 Productive Capacities of Scientific Inquiry

Historical materialism contributes a useful distinction between productive capacities and the relations through which production is organized. In Marx's analysis, productive forces include capacities such as labor power, skill, tools, and means of production, while relations of production concern the social relations within which those capacities are organized [16, 11]. As emphasized in Section 2, the present paper does not map these categories mechanically onto science. It uses them as a structural distinction for separating what a research system is capable of doing from how those capabilities can actually be combined.

The productive capacities of scientific inquiry include more than instruments. They encompass trained researchers, accumulated techniques, computational resources, experimental platforms, measurement systems, datasets, models, software, archives, standards, and forms of tacit expertise. AI systems themselves increasingly belong to this layer. A language model, theorem prover, automated laboratory, or program-search system can expand the set of transformations that a research community is capable of performing. The important point is that the existence of such a capacity does not entail its effective availability for every problem.

Capacities also interact. A powerful model without access to reliable data can have little scientific leverage. A high-resolution instrument without suitable calibration or analysis methods may produce observations that cannot yet support inference. A large archive without metadata or interoperability can remain practically opaque. Scientific productivity therefore depends on configurations of capacities rather than on isolated assets. Historical development can consist in constructing interfaces among capacities that previously existed separately.

This relational view is especially important for AI because contemporary systems can make latent capacities visible. A model may identify that a method developed in one field can be transferred to another, that an overlooked dataset can test a disputed hypothesis, or that a software tool can automate a previously expensive analysis. Such recombination can increase the effective productive capacity of inquiry without creating a wholly new resource. Yet even here historical conditions matter: transfer requires conceptual compatibility, access, validation, and often institutional permission or human expertise. A hypothetical connection generated in text becomes scientifically productive only when the surrounding system can realize it.

The concept of productive capacity also gives a more precise meaning to intermediate research. A project can be valuable because it enlarges the capabilities available to later inquiry. Building a measurement device, establishing a reference dataset, developing a robust protocol, clarifying a classification, or creating a reusable method can change the set of future operations available to a field. The immediate paper or experiment is therefore only one output. The second output is a transformed capacity structure.

This suggests a temporal criterion of scientific value that differs from immediate answer quality. A research action can have modest direct payoff and large developmental significance if it increases the future system's ability to discriminate, measure, coordinate, or reason. The later sections will develop this point as the generativity of intermediate problems. At the present stage, the important conclusion is simply that solvability depends on accumulated capacities, and that research can produce capacities as well as conclusions.

DRAFT

4.5 Relations of Knowledge Production

Productive capacities do not operate in isolation. They are organized through relations of access, authority, ownership, coordination, evaluation, division of labor, and responsibility. Social epistemology provides a broad framework for understanding why such relations matter to knowledge production: criticism, trust, expertise, diversity, and institutional structure can affect what evidence circulates and how claims are evaluated [14]. A historically materialist AI would need to represent these relations when they materially constrain the path of inquiry.

Access relations provide the clearest example. A dataset can exist while remaining unavailable to the researchers or agents capable of analyzing it. An instrument can exist while access is rationed by geography, cost, scheduling, security, or institutional membership. A model can have the abstract capability to perform an analysis while lacking permission to use the relevant data. From the standpoint of scientific reachability, existence and access are different variables. An adequate problem representation must therefore distinguish the resources that exist from the resources that can be legitimately mobilized within the present research arrangement.

Division of labor creates another layer. Modern science often depends on specialization so deep that no individual understands every component of a project. This specialization creates productive power but also coordination problems. Knowledge can become fragmented across disciplines whose vocabularies, incentives, and evidential standards differ. A contradiction can therefore arise in which increasing specialization expands local competence while reducing the ease with which distributed capacities are integrated. The bottleneck is not necessarily insufficient intelligence in any participant. It can lie in the relation among participants.

Evaluation and publication structures influence development in a different way. Metrics, peer-review practices, prestige hierarchies, funding criteria, and publication incentives help allocate scarce attention and resources. They can support quality control and cumulative specialization. They can also make certain forms of work difficult to sustain, especially infrastructure building, replication, negative results, maintenance, or exploratory work whose value appears only later. The point is not that evaluation is inherently distorting. It is that evaluative relations can become part of the causal explanation for why a field develops along one path rather than another.

Ownership and governance matter for similar reasons. Scientific tools and datasets can be public, proprietary, restricted, or jointly governed. Different arrangements alter reuse, verification, interoperability, and the distribution of research opportunities. Ethical governance also establishes boundaries that a responsible AI must not interpret merely as obstacles to optimization. Some constraints protect persons, communities, environments, or legitimate interests. A historically materialist analysis therefore cannot treat every restriction as a contradiction to be removed. It must distinguish conditions that unnecessarily suppress inquiry from conditions that rightly structure permissible inquiry.

This is one reason the framework cannot be reduced to maximizing scientific throughput. Relations of knowledge production contain normative commitments as well as causal effects. The fact that a change would accelerate research does not by itself justify the change. A future autonomous system must reason under constraints of legality, ethics, consent, safety, and distributive consequences. Historical-material analysis can reveal how these relations shape solvability, but it does not automatically determine which transformations ought to be pursued.

For the present paper, the analytical role of these relations is narrower. They explain why productive capacities and effective scientific capabilities can diverge. A field can possess sufficient technical resources in aggregate while lacking the relational organization required to combine them. Conversely, a change in coordination, access, standardization, or governance can unlock research possibilities without any dramatic increase in raw technical capacity. This divergence will later become important for contradiction analysis: some apparently conceptual problems may be reproduced by relational conditions, while some institutional problems may be consequences of deeper epistemic or material limitations.

DRAFT

4.6 Solvability Across Historical States

The preceding distinctions permit a more precise statement of the paper's central object. Scientific solvability is not treated as a timeless property of a problem considered in isolation. It is a relation between a problem and a historically constituted research state. A problem can therefore remain linguistically stable while its solvability changes because the surrounding conditions of inquiry change.

At a schematic level, one can write

$$\text{Solvable}(P, t) = \text{Solvable}(P \mid S_t),$$

where P denotes a problem and S_t denotes the historically available configuration of knowledge, evidence, methods, instruments, productive capacities, and relations of inquiry. The notation should not be read as a binary decision rule. "Solvable" can include degrees of tractability, uncertainty, cost, and evidential adequacy. The important move is to place the historical state inside the problem rather than outside it.

This relational definition immediately explains why scientific planning can require transformation rather than direct attack. If P is not presently researchable from S_t, the relevant next action may be an intervention a_t that produces a new state S_{t+1}. The scientific value of a_t then depends partly on whether it changes the relation between the field and P. A dataset, instrument, method, conceptual distinction, or coordination structure can therefore function as a developmental intervention even when it does not answer P directly.

The same perspective prevents the easy problem from being misunderstood as mere pedagogical decomposition. Ordinary decomposition takes a difficult target and divides it into smaller components whose solutions can later be combined. Historical development can have a stronger structure. Solving the intermediate problem changes the available research system itself. The next problem is posed from a different state. What is learned is not merely another piece of the final answer; new capacities, evidence, categories, or relations have been produced.

This makes research path-dependent in an epistemically significant sense. Different sequences of intermediate interventions can create different future opportunity structures. A field that invests early in measurement infrastructure may later formulate different questions from a field that invests primarily in theory. A community that standardizes data sharing may produce different cumulative evidence from one organized around isolated proprietary datasets. Neither path is guaranteed to be globally superior, and history need not converge. The point is that the path helps constitute the later space of possibility.

For AI, the design implication is conceptual before it is computational. An autonomous scientific system should not assume that every important problem already belongs to the presently actionable search space. It should be able to represent a boundary between what can be attempted now and what requires prior transformation. It should also be able to revise that boundary after its own actions, or the actions of others, change the research state. This is the beginning of historical self-evolution: the agent does not merely update an answer within a fixed problem environment; it participates in altering the conditions that define the environment's scientific possibilities.

The next section makes this idea more explicit by introducing the historically constituted research state as an analytical object. The purpose of that representation will not be to catalogue every feature of a scientific field. It will be to identify the subset of historical conditions that are causally relevant to the present reachability of problems and to the transformations available to inquiry.

5 The Historically Constituted Research State

Section 4 treated solvability as a relation between a scientific problem and the historically available conditions of inquiry. That formulation now requires a more explicit account of the state from which inquiry proceeds. The purpose of such an account is not to build an exhaustive ontology of science. No finite representation could include every fact about a field, institution, community, or technological environment, and an autonomous system that attempted to record everything would lose the distinction between

DRAFT

relevant conditions and incidental background. The narrower task is analytical: to represent those aspects of a research situation whose organization changes what can presently be formulated, investigated, discriminated, established, or transformed. A historically constituted research state is therefore a selective representation of the conditions that bear on scientific reachability.

The section first clarifies the analytical role of a research state and distinguishes it from a mere inventory of resources. It then introduces a relational representation in which knowledge, evidence, methods, material capacities, institutions, and access conditions matter through their organization as well as through their presence. The discussion next adds historical trajectory and path dependence, because a present configuration does not fully explain how it can develop. It then connects the research state to reachable and unreachable spaces of inquiry, treats research action as a state transition, and closes by emphasizing partial observability and the fallibility of any reconstruction. These qualifications are important for the later analysis of contradiction: one cannot identify a structurally central tension without first specifying the state in which that tension is reproduced.

5.1 The Analytical Role of a Research State

A research state is best understood as a problem-relative description of an evolving scientific situation. The adjective "state" can otherwise suggest a static snapshot, as if inquiry could be frozen into a complete list of variables. That is not the intended meaning. The concept instead identifies a temporary configuration from which particular transformations are possible and others are not. It functions in the same broad analytical spirit as a state description in dynamical reasoning: not as a claim that the world is exhausted by the representation, but as a device for asking what changes are reachable from the present configuration.

This problem-relative character matters. The conditions relevant to one scientific question can be largely irrelevant to another. Access to a particular telescope may be decisive for one astronomical problem and immaterial to a mathematical proof. A regulatory framework can determine whether a medical experiment is permissible while having no bearing on a purely computational benchmark. The research state should therefore not be defined as "everything science currently possesses." It is the subset of conditions that bear causally or normatively on the inquiry under consideration.

A useful first approximation is to write

$$S_t = \langle X_t,\, \Gamma_t,\, H_t \rangle,$$

where X_t denotes the presently relevant elements or capacities, Γ_t the relations through which those elements are connected and made available, and H_t the trajectory through which the present configuration has been formed. This notation is deliberately abstract. It does not assume that all scientific fields share the same state variables, nor that each element can be assigned a scalar quantity. Its purpose is to prevent three reductions that would be especially damaging for a historically oriented AI.

The first reduction would identify a state with its stock of knowledge. Established propositions matter, but a field can know what it needs to measure while lacking the instrument with which to measure it. The second reduction would identify a state with its stock of resources. Two communities can possess similar data, tools, and expertise while differing greatly in whether those capacities can be combined because access, standards, authority, or coordination differ. The third reduction would identify the present configuration with its current snapshot. Two apparently similar states can have different future possibilities because they were produced through different paths, contain different tacit dependencies, or reproduce their capacities through different institutions.

The state concept therefore has a functional role. It allows an artificial scientist to ask why a problem occupies its current epistemic position. If a question lies outside present researchability, the system should be able to identify which features of S_t are responsible. If several interventions are available, it should be able to reason about which components and relations each intervention would alter. If a previous intervention has already changed the field, it should not continue planning from an obsolete description. Historical self-evolution begins from this capacity to reconstruct the situation after transformation.

DRAFT

The representation should also distinguish the research state from the AI's internal state. A model may improve its memory, planning, tools, or reasoning policy while the surrounding scientific situation remains unchanged. Conversely, the scientific situation may change because another laboratory releases a dataset, a new instrument comes online, a legal rule changes, or a conceptual distinction becomes widely accepted even if the AI itself is unmodified. In many cases both processes occur together. The distinction is analytically useful precisely because it permits the question of whether a failure should be addressed by changing the agent, changing the surrounding conditions, or changing both.

5.2 Relational Composition of the Research State

The elements gathered in X_t can be grouped provisionally into several families: established and contested knowledge, available evidence, methods and inferential procedures, instruments and computational capacities, skilled human and artificial labor, institutional structures, standards, and legally or ethically relevant permissions. The grouping is not intended as a fixed taxonomy. Different fields may require different decompositions, and the boundaries between categories are themselves historically variable. A statistical method can become embedded in software infrastructure; a database can function simultaneously as evidence, infrastructure, and a coordination standard; an AI model can be a method, an instrument, and a participant in a division of cognitive labor.

For this reason the relational structure Γ_t is at least as important as the inventory X_t. Scientific capacities become effective through relations of compatibility, access, dependence, authority, communication, trust, ownership, validation, and coordination. A dataset and an analytical method may both exist yet fail to compose because the variables are defined differently. Two laboratories may possess complementary expertise yet be unable to collaborate because of legal restrictions or incompatible protocols. A powerful instrument may produce observations whose interpretation depends on calibration standards maintained elsewhere. The state of inquiry is therefore not a bag of resources. It is a structured configuration in which the effect of one element depends on its relations to others.

Effective capability is therefore relational rather than additive. Adding more compute, data, or personnel need not increase what a field can actually do if the bottleneck lies in interoperability, validity, access, or conceptual coordination. The scientific significance of an element depends partly on how it is connected to other elements and on whether those connections can be legitimately mobilized for the problem at hand.

The relational view also prevents a common mistake in autonomous planning: treating all unavailable resources as missing resources. A needed capacity may already exist somewhere in the scientific system but remain disconnected from the present problem. The appropriate intervention would then be relational rather than productive in the narrow sense. Standardizing a schema, negotiating legitimate access, translating between disciplinary vocabularies, or establishing a shared protocol can make an existing capacity usable. Such actions change Γ_t while leaving much of X_t intact.

Relations can also suppress or distort the use of capacities. Evaluation structures may make long-horizon infrastructure work difficult to sustain. Prestige and disciplinary boundaries can affect which evidence receives attention. Proprietary control can limit reproducibility. Conversely, governance and ethical constraints can legitimately restrict actions that would otherwise expand experimental reach. A research-state representation should therefore encode not only causal availability but also permissible availability. A capacity that could technically be mobilized but should not be mobilized under applicable ethical or legal constraints is not simply an unused opportunity waiting for optimization.

The resulting picture is more demanding than ordinary context retrieval. An AI can retrieve that a dataset exists, that a laboratory owns an instrument, or that a method has been published without understanding how those facts jointly constrain the problem. Historical reconstruction requires relations among facts: which method validates which measurement, which institution controls access, which standard makes two datasets comparable, which dependency would fail if a service disappeared, and which forms of coordination would be required to combine distributed capacities. The relevant state is thus organized, not merely accumulated.

DRAFT

5.3 Historical Trajectory and Path Dependence

A present configuration does not contain its own history transparently. This is why H_t appears explicitly in the representation. Historical trajectory matters whenever the same or similar present elements can support different future developments because their dependencies, meanings, or reproductive mechanisms differ. A field's vocabulary, evidential standards, institutional divisions, and technical infrastructure are products of prior selections and accommodations. They can therefore carry constraints that are not visible from a purely synchronic inventory.

The most straightforward case is infrastructural dependence. A method may rely on a chain of standards, software, datasets, calibration practices, and tacit expertise accumulated over decades. Replacing one visible component can be much harder than its present description suggests because other elements have been organized around it. Institutional arrangements exhibit similar effects. A division of labor that once solved a coordination problem can later make interdisciplinary integration difficult, even when all parties agree that integration would be useful. Historical development can thus produce both capabilities and rigidities.

Conceptual systems are path-dependent as well. Scientific terms acquire operational meanings through measurement practices, model classes, exemplary cases, and disputes about classification. A later concept can preserve distinctions inherited from earlier research even when alternative classifications are logically possible. Kuhn's historical account of scientific development is relevant here because it emphasizes that standards, exemplars, and problem significance can evolve with scientific practice rather than against an invariant conceptual background [3, 18]. The present argument does not require the stronger claim that every research state is paradigm-bound. It requires only that the meaning and usefulness of present categories can depend on the sequence through which they were stabilized.

For AI, historical trajectory performs at least three functions. First, it supplies explanation. The system can distinguish a condition that is accidental in the present snapshot from one reproduced by a durable mechanism. Second, it supplies transition information. A state variable that is easy to change in one historical configuration may be deeply embedded in another. Third, it supplies alternatives. Past branches, abandoned methods, negative results, and earlier conceptual distinctions can reveal possible transitions that are invisible if history is compressed into the currently dominant arrangement.

This makes historical memory different from simple archival accumulation. Storing every prior paper or experiment does not by itself represent history. The system needs some account of dependency and transformation: what problem an earlier intervention addressed, what capacity it created, what relation it displaced, what new limitation followed, and which alternatives ceased to be pursued. History becomes analytically useful when it helps explain the present transition structure.

The trajectory term also guards against teleology. If the current state is treated as the necessary culmination of previous research, the AI will tend to interpret earlier alternatives only as failed approximations to the present. A historically self-evolutionary system should instead preserve the possibility that prior branches embodied different tradeoffs or unrealized opportunities. The fact that one path became dominant does not prove that it was the only viable path or that its continuation is now appropriate. Later sections will return to branching and counterfactual histories; at this point the key claim is that a state description should retain enough trajectory information to make alternative development intelligible.

5.4 Research Reachability and Opportunity Structure

Once the research state is represented relationally and historically, the distinction between reachable and unreachable inquiry can be stated more precisely. Let $\mathcal{Q}(S_t)$ denote the set of questions that can be meaningfully pursued from the present state under available methods, evidence, capacities, and permissible interventions. The notation should again be read schematically. Reachability is rarely binary, and different communities can disagree about whether a path is sufficiently mature to count as viable research. What matters is that the set depends on the state.

The same underlying scientific concern can therefore generate different actionable questions at different times. An early state may permit only coarse observation and descriptive classification. A later

DRAFT

state may support controlled intervention, mechanistic discrimination, or fine-grained causal analysis. The difference need not arise because researchers have become more intelligent in an abstract sense. The opportunity structure of inquiry has changed. New instruments, standards, theories, datasets, or organizational forms make different transformations possible.

This gives a clearer meaning to the boundary introduced in the preceding section among conceptual possibility, researchability, and establishability. Conceptually formulable questions can exist outside $\mathcal{Q}(S_t)$ if no available path can yet produce discriminating evidence. Some questions inside the researchable region can remain outside present establishability because evidence is too weak or standards of inference are not yet satisfied. The boundaries are dynamic surfaces of a research state, not permanent classifications attached to propositions.

A useful consequence follows for research planning. The objective need not be to select the most important question in the abstract. The relevant question can instead be which reachable intervention most productively changes the future opportunity structure. An action a_t may have limited immediate scientific payoff while expanding the set of questions available after the transition. If

$$S_t \xrightarrow{a_t} S_{t+1},$$

then one important property of the action is the relation between $\mathcal{Q}(S_t)$ and $\mathcal{Q}(S_{t+1})$. A new instrument, dataset, protocol, conceptual distinction, or coordination mechanism can enlarge or reshape the region of reachable inquiry even if it does not directly answer the motivating problem.

Expansion is not the only possible change. A transition can also close paths. Ethical findings can appropriately rule out an experimental procedure. Better evidence can show that a once-promising approach is based on an invalid assumption. Loss of infrastructure can reduce researchability. Standardization can improve comparability while making some forms of exploratory heterogeneity harder to maintain. The opportunity structure should therefore not be imagined as a quantity that simply increases with time. Historical development changes its shape.

This point is especially important for self-evolutionary AI because optimization language encourages monotonic interpretations. If every transition is scored by a single notion of progress, the system may overlook tradeoffs among different future capacities. A historically oriented planner should instead be able to represent that one intervention opens some regions of inquiry while narrowing others. The later requirement of plural developmental paths follows from this structure: preserving several viable branches can be rational when the long-term scientific consequences of early transitions remain uncertain.

5.5 State Transition and Historical Intervention

Treating research as movement among states changes the meaning of scientific action. A research action is not only an attempt to extract an answer from a fixed environment. It can modify the environment from which later inquiry proceeds. Experiments produce evidence, but they can also reveal measurement limitations. Methods solve tasks, but they can also create reusable capacities. Conceptual work can resolve ambiguity, but it can also generate new categories that reorganize subsequent observation. Institutional interventions can alter access or coordination. The output of research is therefore partly a changed state.

At the most schematic level, the transition can be written as

$$S_{t+1} = \mathcal{T}(S_t, a_t, o_t),$$

where a_t is an intervention and o_t denotes outcomes that are not fully controlled by the agent. The inclusion of o_t is philosophically important. Scientific development is not a deterministic planning problem in which the AI chooses the next state directly. Experiments surprise, instruments fail, other actors intervene, institutions change, and external events alter available resources. The transition function therefore represents conditioned development, not command over history.

An autonomous system must consequently distinguish intended transformation from realized transformation. It may choose an intervention because it expects a dataset to resolve a bottleneck, only to discover that the data expose a more fundamental problem in the measurement model. That result should

DRAFT

not be treated simply as failure. It changes S_t by altering what is known about the obstacle. Historical self-evolution requires the AI to reconstruct the state after observation rather than forcing new evidence into the assumptions that motivated the intervention.

The action space is heterogeneous. Some interventions operate on evidence; others on methods, instruments, concepts, standards, or relations of coordination. Still others operate on the AI itself. This heterogeneity is exactly why the research-state representation is useful. It gives the planner a common question across otherwise different actions: which part of the historically constituted state is being transformed, and what new paths does that transformation make available or unavailable?

The framework also makes room for indirect intervention. If a scientific problem is blocked by an absent measurement capacity, building the capacity may be more rational than generating additional hypotheses. If the bottleneck is incompatible data standards, coordination may dominate computation. If the central difficulty is theoretical underdetermination, collecting more of the same evidence may have little value until a discriminating experiment is designed. The next action is selected through diagnosis of the state rather than through a fixed preference for any one type of scientific activity.

This perspective prepares the transition to contradiction analysis. A research state normally contains many limitations at once. Some are relatively independent, while others are linked through common dependencies. Some disappear when another condition changes; some persist because the system continually reproduces them. The central task is therefore not merely to list everything that is missing from S_t, but to understand the structure among its tensions. The next section will develop that problem.

5.6 Partial Observability and Fallible Reconstruction

No scientific agent has complete access to the research state it attempts to model. Evidence can be unpublished, tacit expertise difficult to articulate, institutional incentives opaque, dependencies undocumented, and future reactions unknowable. Even when information is publicly available, the relevance of a condition to the present problem may be uncertain. The historically constituted research state is therefore never given directly. It must be inferred.

This qualification matters because the language of state representation can otherwise create an illusion of omniscience. The AI does not stand outside history with a complete map of productive forces, relations, contradictions, and future transitions. It is one participant situated within the system. Its own tools, training data, access permissions, evaluative assumptions, and institutional embedding are themselves parts of the conditions it seeks to reconstruct. Historical self-evolution consequently requires reflexivity: the system should be able to represent uncertainty about its model of the state and recognize when its own position limits what it can observe.

A practical representation would therefore need confidence, provenance, and contestability rather than a single authoritative state estimate. Competing reconstructions may be appropriate when evidence is incomplete. Human researchers and affected institutions can possess local knowledge unavailable to the AI. A relation inferred from publication patterns may have a different explanation from the one suggested by the data. The system should preserve such alternatives long enough for further evidence to discriminate among them.

Fallibility also constrains intervention. A mistaken diagnosis of the research state can lead an autonomous planner to transform the wrong condition. If it misclassifies a conceptual limitation as a resource shortage, it may scale data collection without resolving the underlying ambiguity. If it misclassifies a legitimate ethical boundary as an institutional bottleneck, it may recommend an impermissible intervention. If it interprets a historically contingent arrangement as necessary, it may reproduce avoidable lock-in. State reconstruction is therefore an epistemic task subject to the same requirements of evidence and revision as other scientific claims.

The result is a deliberately modest conception of S_t. It is neither a complete world model nor a fixed vector supplied to an optimizer. It is a revisable, problem-relative reconstruction of the historically relevant conditions of inquiry. Its value lies in making explicit what ordinary autonomous search often leaves implicit: scientific problems are approached from structured states, those states have histories, interventions transform them, and the boundary of reachable inquiry moves with those transformations.

DRAFT

With this object in place, the next question becomes sharper. Among the many tensions present in a research state, which ones organize the development of the others, and how can their relative structural importance change through time?

6 Contradiction Structures in Scientific Development

The preceding section introduced the historically constituted research state as a revisable representation of the conditions from which inquiry proceeds. A research state, however, is not only a collection of capacities and relations. It also contains tensions among requirements, practices, resources, standards, and developmental tendencies. Some of these tensions are local obstacles. Others are reproduced by the organization of the research state itself and condition several downstream difficulties at once. If historically self-evolutionary problem solving is to proceed by diagnosis rather than by undirected search, it needs a way to distinguish these cases. The present section develops such an account.

The purpose is not to impose a universal dialectical vocabulary on every scientific difficulty. A failed experiment, a missing dataset, a logical inconsistency, an ethical prohibition, and an institutional conflict are not automatically instances of the same structure. The term *contradiction* is used here in the qualified sense established in Section 2: a structured tension or mutually constraining relation that is generated or sustained by the present organization of inquiry and that matters to its subsequent development. The analysis therefore begins by distinguishing contradiction from mere disagreement or shortage. It then examines dependencies among coexisting tensions, clarifies structural centrality, considers mechanisms through which contradictions are reproduced, and explains why contradiction hierarchies must be reconstructed as research states change. The section closes with the epistemic limits of contradiction diagnosis, preparing the later discussion of dialectical transformation.

6.1 Contradiction as a Relational Property of a Research State

A contradiction relevant to scientific development should not be represented as an isolated object that simply exists inside a field. It is a relation among elements of a research state. The same instrument, institution, method, or standard can participate in different contradictions depending on how it is connected to other requirements. A scarcity of observations, for example, is not by itself a contradiction. It becomes part of a developmental contradiction when the research programme simultaneously requires a level or kind of evidential discrimination that the available observational regime cannot supply, while the programme continues to depend on that regime.

This relational character can be represented schematically. Let the research state at time t be S_t, as introduced in Section 5. A contradiction structure associated with that state can be written as

$$\mathcal{C}(S_t) = \{C_1, C_2, \ldots, C_n\},$$

where each C_i represents a problem-relevant tension whose terms and reproduction conditions are specified with reference to S_t. The notation is deliberately weak. It does not assume that contradictions are discrete, mutually exclusive, or completely enumerable. It only makes explicit that contradiction claims are indexed to a reconstructed historical state rather than treated as timeless properties of a scientific domain.

The two sides of a contradiction should likewise not be mistaken for two propositions in a debate. Consider a research programme that seeks increasingly fine-grained mechanistic explanation while relying on measurements whose spatial or temporal resolution cannot distinguish the mechanisms under consideration. The relevant relation is not simply "fine-grained explanation versus coarse measurement" as two verbal positions. It is a dependency: the programme's explanatory ambition generates evidential demands that its material measurement regime cannot satisfy. The contradiction is located in the organization of the practice, not in linguistic opposition between sentences.

This point becomes more important when social and institutional conditions are involved. A field may require broad data sharing for replication while operating under legitimate privacy, security, consent,

DRAFT

or ownership constraints. The tension is not properly understood by declaring one side progressive and the other obstructive. Both requirements can be justified, and the developmental problem lies in the relation between them. A new governance mechanism, privacy-preserving technique, or institutional arrangement might transform that relation. The contradiction therefore identifies a structure requiring mediation, not an enemy to be eliminated.

Contradictions can also span analytical levels. A conceptual demand can depend on a material capacity; an institutional rule can shape which evidence becomes available; a technical standard can reinforce a division of labour; an evaluative norm can reward practices that reproduce a methodological bottleneck. For this reason, the contradiction structure of a research state cannot be reconstructed solely from published arguments. It may require evidence about instruments, workflows, access, incentives, coordination, and the historical development of the field. Historical-material analysis enters precisely at this point: tensions in knowledge production are situated in organized practices rather than reduced to conflicts among ideas.

6.2 Distinguishing Developmental Contradictions from Local Obstacles

The analytical usefulness of contradiction depends on restraint. If every unsolved problem or inconvenience is called a contradiction, the concept loses explanatory value. A historically self-evolutionary AI therefore needs criteria for distinguishing developmental contradictions from other kinds of difficulty.

A *local obstacle* can impede research without being generated by the organizing relations of the research state. A temporary equipment failure, an accidental loss of a file, or a one-time scheduling conflict may be costly, but resolving it need not reorganize the developmental structure of inquiry. A *shortage* can likewise be important without being contradictory. If a laboratory requires more storage capacity and additional storage can be acquired without changing any other relevant condition, the problem may be straightforward scarcity. A *logical inconsistency* is different again: incompatible propositions can require revision, but their inconsistency does not by itself establish a historically generative tension in the organization of research. Finally, a *disagreement* among scientists can persist because evidence is incomplete or values differ, yet the disagreement becomes a developmental contradiction only when its persistence is connected to the structure through which inquiry is reproduced or constrained.

A stronger diagnosis requires at least three features. First, the tension should be *relationally grounded*: its opposed or mutually constraining tendencies should arise from identifiable elements and relations within the reconstructed state. Second, it should be *recurrently reproduced* or structurally persistent rather than merely episodic. Third, it should possess *developmental relevance*: altering the relation would change what kinds of inquiry, coordination, evidence, or evaluation become possible. These features do not form a mechanical checklist, but they help prevent the language of contradiction from becoming a rhetorical label.

The distinction is especially important for AI because language models are highly capable of generating oppositions on demand. Given any scientific proposal, an LLM can usually produce a countervailing consideration. This facility creates a risk of pseudo-dialectics: the appearance of depth through balanced pairs that have no demonstrated role in the actual development of the research situation. A system may describe "accuracy versus interpretability," "innovation versus safety," or "speed versus rigor" even when one of those pairs is not materially relevant to the case under examination. Dialectical analysis requires evidence that the relation is immanent to the state being reconstructed.

The same caution applies to familiar tradeoffs. A tradeoff describes a situation in which improving one objective tends to worsen another under a given set of constraints. Some tradeoffs can be dialectically significant, but they need not remain permanent. A tradeoff can disappear after the state changes: a new instrument may improve both speed and precision; a new governance arrangement may improve both access and privacy; a conceptual distinction may dissolve an apparent opposition. Treating every tradeoff as an immutable law would therefore obscure historical transformation. The contradiction concept is useful precisely when it directs attention to the conditions that produce the apparent tradeoff and asks whether those conditions themselves can change.

DRAFT

This also distinguishes contradiction analysis from generic risk assessment. Risk assessment asks what adverse outcomes may follow from an action. Contradiction analysis asks how the existing organization of inquiry generates mutually constraining tendencies and how those relations condition the path of development. Both can be necessary, but they answer different questions. The latter is specifically concerned with the historical structure of the problem situation.

6.3 Dependency and Propagation across Coexisting Tensions

Research states normally contain multiple contradictions at once. Their importance cannot be understood by inspecting each one independently because tensions can depend on, amplify, mask, or reproduce one another. The contradiction structure is therefore better understood as a dependency pattern than as a flat list.

Suppose a field experiences weak replication, fragmented datasets, incompatible measurement standards, and strong pressure for rapid publication. These difficulties may appear as four separate problems. Historical reconstruction can reveal a different structure. Fragmented datasets may partly result from incompatible standards; weak replication may be intensified by both fragmentation and publication incentives; pressure for speed may discourage investments in shared infrastructure that would otherwise reduce fragmentation. The scientific difficulty is then distributed across a set of relations. Intervening on one tension can alter several others even when they are not directly targeted.

This dependency can be represented minimally by a relation among contradictions. If $C_i \rightsquigarrow C_j$ denotes evidence that the reproduction or severity of C_j depends materially on C_i under the present state, then the structure

$$C_i \rightsquigarrow \{C_j, C_k, C_\ell\}$$

expresses a stronger developmental role than a collection of independent tensions. The arrow should not be read as a universal causal law. It records a historically situated dependency claim that remains open to revision. In many cases the relation will be bidirectional or mediated by additional conditions.

Propagation can occur through several mechanisms. A material bottleneck can constrain many methods that depend on the same instrument. An institutional rule can shape multiple laboratories simultaneously. A conceptual ambiguity can contaminate measurement, annotation, and evaluation practices across a field. A funding structure can favour short-horizon outputs and thereby weaken infrastructure, replication, and cumulative data stewardship. Conversely, a local technical failure can remain isolated even if it attracts significant attention. Structural analysis therefore asks not only how severe a contradiction appears, but how widely its effects propagate through the relations constituting the research state.

Dependency also explains why symptom-oriented intervention can produce disappointing results. If several downstream tensions are generated by a shared upstream relation, addressing them separately can consume substantial effort while leaving the reproducing structure intact. A field may repeatedly improve reporting guidelines, for example, while the material incentives and infrastructure that produce non-reproducible workflows remain unchanged. The point is not that upstream intervention is always preferable. Some upstream conditions may be immutable, ethically protected, or prohibitively difficult to change. The analytical claim is narrower: the dependency structure should be reconstructed before the developmental role of each contradiction is assessed.

The dependency pattern can also reveal hidden contradictions. A recurring cluster of apparently unrelated failures may indicate a common relation that has not yet been represented. In this sense contradiction analysis can be abductive. The AI observes that several problems co-vary or reappear after local fixes and hypothesizes a deeper structural tension. Such a hypothesis must then be tested against historical and material evidence. This is one way in which a self-evolutionary system could move beyond surface problem lists toward explanations of why a research state repeatedly produces similar limits.

6.4 Structural Centrality and Derivative Tensions

The distinction between structurally central and derivative contradictions introduced philosophically in Section 2 can now be specified in relation to the research state. Structural centrality concerns the role

DRAFT

a contradiction plays in organizing the dependency pattern of present inquiry. It does not mean that the contradiction is morally more important, more visible, more difficult, or more fundamental in an ontological sense. It means that changes in the relation are expected to reorganize a comparatively large or consequential part of the reachable research structure.

Several forms of centrality can matter. A contradiction can be *dependency-central* when many other tensions rely on it. It can be *reproduction-central* when the routines of the research system continually regenerate it. It can be *reachability-central* when transforming it would open or close a substantial region of future inquiry. It can be *coordination-central* when it links otherwise separate actors or infrastructures. These forms need not coincide. A highly connected contradiction can have little effect on research reachability, while a narrowly localized measurement limit can block an entire class of decisive experiments.

For this reason, a historically self-evolutionary AI should resist reducing structural centrality to a single numerical ranking. Graph-theoretic measures, causal estimates, or resource analyses may eventually contribute to diagnosis, but the philosophical concept is multidimensional. A contradiction can be central under one developmental objective and secondary under another. The importance of a privacy–access tension, for example, may differ depending on whether the immediate aim is exploratory model development, clinical validation, or cross-institutional replication. Centrality is indexed to the state, the problem, and the anticipated pathways of transformation.

Derivative contradictions are correspondingly those whose persistence depends substantially on other tensions or whose transformation leaves the broader structure mostly intact. The term *derivative* should not be read as dismissive. A derivative contradiction can cause serious practical or ethical harm and can require immediate intervention. The distinction concerns explanatory and developmental dependence. A downstream replication difficulty can be derivative of incompatible measurement standards while still deserving urgent attention because unreliable results affect patients, policy, or subsequent research.

A useful diagnostic test is counterfactual. If C_i were substantially transformed while the rest of the state remained as similar as possible, which other tensions would plausibly change? Conversely, if a downstream tension C_j were locally repaired, would C_i tend to regenerate it or produce an equivalent obstacle elsewhere? Such counterfactual reasoning does not prove centrality, but it forces the system to articulate why a hierarchy has been inferred. It also provides a basis for comparing rival reconstructions of the contradiction structure.

Structural centrality should therefore be accompanied by an explanatory trace rather than a label. An AI that announces "the central contradiction is X" without identifying dependencies, evidence, and uncertainty would merely reproduce the authority problem that historical analysis is supposed to illuminate. The diagnosis should show which observations support the claimed hierarchy, what alternative hierarchies remain plausible, and what future evidence could revise the ordering.

6.5 Reproduction and Persistence of Contradictions

Some contradictions matter historically because the system repeatedly reproduces them. Their persistence is not adequately explained by the fact that no one has yet found a clever solution. The relations that organize inquiry can recreate the same tension after local interventions.

Reproduction can be material. A scientific programme may depend on an instrument whose physical limitations repeatedly constrain measurement, so every new hypothesis eventually encounters the same boundary. It can be institutional: evaluation rules may reward outputs that discourage the maintenance of shared infrastructure, causing coordination problems to reappear despite repeated calls for collaboration. It can be conceptual: an unstable category can generate inconsistent annotations across successive datasets. It can also be relational: access arrangements can concentrate capabilities in a way that repeatedly produces asymmetries between data holders and external researchers.

Historical persistence therefore requires attention to mechanism. The statement that a field "has always struggled with replication" is descriptive but insufficient. A stronger account asks what practices, incentives, infrastructures, or conceptual commitments cause replication difficulties to recur. The answer can change through time. What was once primarily a measurement problem can later become a

DRAFT

coordination problem after instrumentation improves. The continuity of the label should not obscure discontinuity in the reproducing mechanism.

The concept of reproduction also prevents a simplistic equation between duration and centrality. Some contradictions persist because they are peripheral and receive little attention. Others persist because they are deeply embedded in the organization of inquiry. A long-lived problem is not automatically a structurally central one. Conversely, a newly emergent contradiction can become central rapidly if a technological or institutional transition reorganizes many dependencies around it.

Reproduction can also involve the AI itself. Once autonomous systems participate in literature review, experiment selection, code generation, evaluation, and publication, their own models and incentives become part of S_t. An AI trained predominantly on historically dominant methods may repeatedly recommend those methods, increasing their visibility and further reinforcing the training distribution available to future systems. A ranking system can privilege easily measurable outputs and thereby reshape what researchers choose to produce. Historically self-evolutionary AI must therefore be capable of diagnosing contradictions in which it is not an external observer but one of the mechanisms of reproduction.

This reflexive case is particularly important because self-improving systems can amplify persistence. If an agent evaluates future modifications using criteria inherited from the current state, it can optimize the mechanisms that reproduce the current contradiction. Improvement relative to a local evaluator can therefore deepen historical lock-in. A contradiction analysis that includes reproduction mechanisms provides one way to detect this possibility before recursive optimization is treated as synonymous with development.

6.6 Temporal Reordering and Emergent Contradictions

Contradiction structures are indexed to historical states because their hierarchy can change. A tension that organizes inquiry at t can become secondary after an intervention, while another relation that was previously weak or invisible becomes decisive at $t + 1$. Development therefore cannot be represented adequately by solving a fixed list of contradictions in a predetermined order.

The clearest examples arise when successful research changes its own bottlenecks. Scarce data can initially constrain a field. New sensing infrastructure, large-scale collection, or automated experimentation may transform that condition. The resulting state can then be constrained by interpretation, curation, provenance, or validation rather than scarcity. Similarly, expensive computation can once be the central limit, only for abundant computation to produce new problems of energy use, evaluation, or uncontrolled model complexity. A contradiction can disappear, change form, or give rise to a new tension through the very process that resolved its earlier configuration.

This temporal reordering has two consequences. First, contradiction diagnosis must be repeated after significant state transitions. An AI should not carry forward the previous hierarchy as a stable property of the domain. Second, developmental success cannot be defined as the progressive elimination of contradiction. A more capable research state can contain new contradictions precisely because it supports activities that were previously impossible. The emergence of a new bottleneck can therefore indicate transformed capacity rather than failure.

The relation among contradiction structures can be written schematically as

$$\mathcal{C}(S_t) \longrightarrow \mathcal{C}(S_{t+1}),$$

where the arrow denotes historical transformation rather than simple subtraction. Some elements can persist, some can disappear, some can change their dependencies, and new ones can emerge. The transition can also reorder centrality even when the set of named tensions remains similar.

This view guards against teleology. There is no requirement that successive contradiction structures become simpler, more unified, or closer to a predetermined final form. A transition can resolve one tension while creating several others; a field can become more capable and more complex at the same time. Historical development can branch, regress, or enter a locally stable configuration that later proves restrictive. The analytical purpose of contradiction hierarchy is therefore diagnostic, not prophetic.

DRAFT

For self-evolutionary AI, temporal reordering implies that a successful intervention must be followed by renewed observation and reconstruction. The system should ask what changed in the research state, which dependencies were altered, and which tensions now organize the reachable frontier. This recursive movement from state reconstruction to contradiction diagnosis and back again supplies the bridge to the next section, where the paper turns from identifying contradiction structures to considering their transformation.

6.7 Fallibility and Contestability in Contradiction Diagnosis

Contradiction analysis is itself a scientific and interpretive activity. The structure $\mathcal{C}(S_t)$ is not directly observable, and claims about centrality depend on the state reconstruction from which they are inferred. A historically self-evolutionary AI should therefore represent contradiction diagnosis as fallible rather than as authoritative classification.

Several uncertainties can enter the analysis. The system may omit a relevant actor or resource from S_t. It may mistake correlation among tensions for developmental dependence. A relation that appears to reproduce a bottleneck may be only a symptom of an unobserved condition. Historical evidence can support multiple interpretations of why a field followed a particular path. Institutional actors can also disagree legitimately about which constraints are central because they occupy different positions in the same research system.

Contestability is therefore a requirement rather than a weakness. A contradiction diagnosis should identify the evidence on which it depends, distinguish observed relations from inferred mechanisms, retain plausible alternative reconstructions, and specify what observations would change the hierarchy. Human researchers should be able to challenge both the representation of the state and the inferred dependencies among tensions. The aim is not to replace scientific disagreement with an automated declaration of what the "real contradiction" is.

This requirement also has an ethical dimension. A system that labels an institutional safeguard, consent requirement, or marginalized perspective as a "secondary contradiction" could misuse developmental language to discount legitimate constraints. Structural centrality does not authorize normative priority. An analytically central relation can be ethically non-negotiable; a derivative tension can demand immediate protection. The contradiction hierarchy developed here concerns the organization of scientific development, not a ranking of whose interests matter.

The same distinction should govern later intervention. Knowing that a relation constrains many research paths does not imply that it ought to be removed. Some constraints constitute the conditions of legitimate inquiry. Historically self-evolutionary reasoning therefore needs both developmental diagnosis and independent normative evaluation. The present paper focuses primarily on the former, while preserving this separation so that dialectical language does not become a license for unconstrained optimization.

The resulting conception of contradiction is deliberately relational, historical, and modest. Contradictions are reconstructed from a problem-relative research state; they are distinguished from local obstacles and arbitrary oppositions; they form dependency structures in which some tensions can be structurally central and others derivative; they persist through identifiable mechanisms of reproduction; and their hierarchy changes as the research state changes. Because each of these claims is fallible, the reconstruction must remain contestable. With this diagnostic structure in place, the next step is to examine transformation: how an intervention can negate, preserve, reorganize, or displace elements of a research state, and how such transformations generate new conditions of scientific problem solving.

7 Dialectical Transformation of Research States

Section 6 treated contradiction as a relational and historical property of a research state. The present section turns from diagnosis to transformation. The central question is not how an artificial system can

DRAFT

eliminate every tension it detects, but how it can distinguish interventions that merely suppress symptoms from interventions that reorganize the conditions under which those symptoms are produced. This distinction is important because scientific development rarely proceeds by removing one isolated obstacle while leaving the rest of the problem situation unchanged. A new instrument can alter evidential standards; a new dataset can expose weaknesses in an existing taxonomy; a change in institutional coordination can make previously impossible experiments routine; a theoretical reformulation can transform what counts as a relevant observation. The object of intervention is therefore the research state itself, including the relations through which problems, methods, and evaluative criteria are reproduced.

The term *dialectical transformation* is used here in a deliberately restricted sense. It does not imply that every scientific transition follows a single Hegelian sequence, nor that contradiction guarantees progress. The concept instead highlights four features developed in Section 2: a tension can arise from the internal organization of a practice; an intervention can negate some aspects of that organization while preserving others; the transformed state can change the meaning and role of inherited elements; and the transformation can generate new contradictions rather than terminate development [17]. For historically self-evolutionary AI, the practical implication is that problem solving must be represented as state transformation followed by renewed diagnosis. The system acts, observes what changed, reconstructs the resulting state, and asks again which tensions now structure further inquiry.

7.1 From Contradiction Diagnosis to Developmental Intervention

A contradiction diagnosis does not yet determine an intervention. Knowing that a field depends on high-resolution evidence while relying on low-resolution measurement does not specify whether the appropriate next step is to redesign an instrument, reformulate the theoretical question, combine heterogeneous data sources, narrow the inferential claim, or reorganize the research programme around an intermediate target. Several interventions can respond to the same tension by altering different parts of the research state. The developmental task is therefore comparative: an agent must reason about what each intervention changes and which future states become accessible as a result.

Using the transition notation introduced in Section 5, an intervention a_t participates in a transition

$$S_{t+1} = \mathcal{T}(S_t, a_t, o_t),$$

where o_t represents observations, consequences, or feedback generated through the intervention. The expression should not be interpreted as a deterministic update rule. Scientific interventions have uncertain outcomes, and their consequences can extend beyond what the actor intended. The notation serves only to make one conceptual point explicit: the value of a_t cannot be assessed solely by inspecting the immediate output of the action. It must also be assessed through the state that the action helps to produce.

This is the first major difference between developmental intervention and ordinary answer selection. If a system compares two hypotheses only by predictive performance on a fixed benchmark, the benchmark supplies a relatively stable evaluative frame. If it compares two research interventions, however, one may improve prediction within the existing frame while another changes the evidential infrastructure through which later hypotheses can be tested. Their consequences occupy different temporal horizons. A historically oriented system must therefore distinguish an intervention's immediate result from its state-transforming effect.

This distinction also prevents structural centrality from becoming a simplistic instruction to "attack the biggest bottleneck." A structurally central contradiction may be physically unchangeable, ethically protected, institutionally slow to transform, or best approached indirectly. A derivative tension may be the most feasible point of intervention because changing it alters the reproduction mechanism of the central relation. Developmental intervention is therefore not equivalent to choosing the most central contradiction and removing it. It requires a theory of leverage that is sensitive to mediation, feasibility, legitimacy, and temporal sequence.

For example, suppose a research programme is constrained by the relation between highly complex models and weakly discriminating evidence. Directly reducing model complexity may sacrifice explanatory capacities that the programme legitimately needs, while immediately building a new instrument may

DRAFT

be financially impossible. A more modest intervention could establish a shared benchmark designed specifically to reveal which model distinctions are empirically consequential. Such a benchmark does not dissolve the original contradiction, but it can reorganize the relation between theory and evidence by clarifying where additional measurement capacity would have the greatest value. The intervention is developmental because it changes what can be rationally attempted next.

7.2 Negation, Preservation, and Reorganization

The idea of sublation developed earlier becomes operationally useful when interventions are analyzed according to what they negate, preserve, and reorganize. A historical transformation is rarely well described as replacement of an old state by an entirely new one. Scientific practices accumulate instruments, conventions, datasets, mathematical formalisms, standards, and institutional responsibilities that cannot simply be discarded each time a limitation is identified. Some are still epistemically valuable; others remain necessary for comparability, safety, or coordination. The problem is therefore selective transformation.

An intervention can *negate* an element in several senses. It can invalidate a hypothesis, remove a procedural requirement, abandon a classification, relax a constraint, or demonstrate that a previous distinction no longer organizes the phenomenon adequately. Negation need not mean deletion. A superseded model can cease to function as the general explanatory framework while remaining useful within a restricted domain. A dataset encoded under an obsolete ontology can remain informative if its provenance is preserved. A failed method can remain historically important because it identifies conditions under which a class of approaches does not work.

Preservation is equally selective. A system should preserve an element because its epistemic, methodological, historical, or normative role remains relevant in the transformed state, not because historical continuity is valuable in itself. This requirement is particularly important for AI systems capable of rewriting their own tools or workflows. Performance-oriented self-modification creates pressure to discard apparently inferior components. Historical self-evolution instead requires enough memory and provenance to determine whether an apparently inferior component performs a role that the current evaluator does not capture. A slower procedure may supply interpretability; an older instrument may provide calibration continuity; a minority methodology may explore regions excluded by the dominant paradigm.

Reorganization is the distinctive third operation. The same element can persist while its relation to the rest of the research state changes. A conceptual distinction can move from foundational principle to local approximation. A laboratory technique can move from primary measurement to validation. Human expertise can shift from routine execution to anomaly interpretation after automation. A centralized repository can become infrastructure for distributed experimentation. These cases are not well described by the binary choice between preserving and eliminating an element. The transformed state assigns inherited components new positions and functions.

For historically self-evolutionary AI, this suggests that intervention planning should include an explicit inheritance analysis. Before a major change, the system should ask which dependencies rely on the element being transformed, what functions must survive, which historical records are required to interpret later results, and which alternative branches should remain recoverable. Such analysis is not conservatism. It is a way of preventing optimization from destroying the very conditions through which future development can be evaluated.

The philosophical point is also methodological. Dialectical transformation cannot be reduced to the familiar prompt pattern of thesis, antithesis, and synthesis. A generated synthesis may combine phrases from opposed positions while leaving the underlying research organization untouched. A genuine transformation, in the limited sense used here, must explain how the proposed intervention responds to a tension generated by the existing state, which elements of the state it changes, which it retains, and how their relations differ afterward. The criterion is historical and structural rather than rhetorical.

DRAFT

7.3 Mediation and Indirect Transformation

Many important scientific contradictions cannot be transformed by acting directly on either side. Their relation is mediated by technical, conceptual, institutional, or social arrangements. The concept of mediation is therefore central to long-horizon problem solving. It shifts attention from choosing between opposed requirements to constructing conditions under which their relation can change.

Consider a tension between broad data access and protection of sensitive information. Treating the problem as a binary conflict invites either unrestricted sharing or restrictive closure. In practice, the relation can be transformed through mediating arrangements such as trusted research environments, privacy-preserving computation, tiered access, governance procedures, standardized consent, or distributed analysis. None of these mechanisms is universally adequate, and each introduces its own costs and risks. Their importance lies in changing the form of the relation. The developmental question becomes which mediation reorganizes the research state in a way that expands legitimate inquiry while preserving protections that should remain binding.

Mediation can also be conceptual. A field facing disagreement between two apparently incompatible models may develop an intermediate representation that identifies the domain in which each model is informative. The result need not be a synthesis into one larger theory. It may instead be a map of regimes, scales, or observational conditions under which different representations apply. The contradiction is transformed because the earlier opposition no longer has the same structure. What appeared to be competing universal claims becomes a differentiated problem of scope.

Institutional mediation is equally important. Some scientific bottlenecks arise because capabilities exist but are distributed across organizations that cannot coordinate effectively. The missing object is not another hypothesis but an interface: common standards, shared infrastructure, legal agreements, funding arrangements, governance procedures, or roles that make cooperation durable. Historical materialism is relevant here because the productive capacity of science depends on relations through which resources and labour are organized, not only on the amount of knowledge available [16, 14]. A historically oriented AI must therefore be able to treat institutional design as part of problem solving when institutional relations are constitutive of the research state.

The importance of mediation also clarifies why small interventions can matter disproportionately. An intermediate protocol, translation layer, benchmark, or classification can connect capacities that previously could not interact. Its immediate scientific result may be modest, yet its relational effect can be large. Such cases prepare the argument developed in the next section: the generative value of research often lies in the future possibilities an intervention creates rather than in the immediate magnitude of its conclusion.

For AI, mediation requires a broader action vocabulary than answer generation. A system concerned with historical development may need to propose experiments, datasets, standards, interfaces, institutional arrangements, validation procedures, or sequences of smaller research tasks. The proposal need not be executed autonomously. The philosophical requirement is more basic: the system must recognize that some contradictions are transformed by constructing new relations among existing elements rather than by selecting one side of an opposition.

7.4 Intervention as Transformation of Reachability

Section 4 distinguished conceptual possibility, present researchability, and present establishability. Dialectical transformation connects those distinctions dynamically. An intervention matters historically when it changes the set of questions that can be responsibly investigated or established from the resulting state.

Let $\mathcal{Q}(S_t)$ denote, schematically, the set of questions or research moves reachable from state S_t. After an intervention,

$$\mathcal{Q}(S_{t+1}) \neq \mathcal{Q}(S_t)$$

may hold even when the intervention does not directly answer the long-term question that motivated it. A new measurement technology can make a previously speculative distinction testable. A shared dataset

DRAFT

can make comparative evaluation possible. A negative result can eliminate an entire family of approaches and redirect resources toward alternatives. A governance arrangement can make multi-institutional evidence collection feasible. The historical significance of the intervention lies in the changed opportunity structure.

This perspective suggests a different way to compare candidate actions. Conventional optimization asks which action produces the best immediate value under a given evaluator. Developmental reasoning asks what research states the action makes reachable, what capacities it creates or destroys, and what dependencies it establishes. Two interventions with similar short-term results can differ radically in their effect on future inquiry. One may produce a high-performing but opaque local solution; another may produce a slightly weaker result while creating a reproducible method, an interpretable dataset, and a platform for subsequent work. The second can have greater developmental value even if the first scores better on the current task.

The idea should not be converted too quickly into a scalar objective such as "maximize future reachability." More options are not automatically better. Some newly reachable actions are unethical, scientifically unproductive, or resource-intensive; some forms of constraint are constitutive of legitimate research. Moreover, increasing the number of available paths can reduce coordination or make validation more difficult. Reachability is therefore an analytical object, not a sufficient normative criterion. A historically self-evolutionary system needs to represent how the option structure changes before independent epistemic and ethical evaluation can determine which changes are desirable.

The same caution applies to irreversible transformations. Large infrastructures, standards, and institutional reforms can create strong path dependence. An intervention that expands one region of the research space can make another region expensive or inaccessible. Developmental comparison should therefore include reversibility and option preservation where relevant. A modest intervention that keeps several trajectories open may sometimes be preferable to a dramatic transformation that commits the field to one uncertain path. This is one reason why historical memory and branching alternatives will remain important later in the paper.

7.5 Emergent Contradictions after Transformation

A transformed research state is not a contradiction-free endpoint. Interventions alter relations, and altered relations generate new tensions. This feature is essential if historical self-evolution is to be distinguished from a problem-solving loop that assumes convergence toward a final stable optimum.

Suppose a field develops an instrument that resolves the evidential scarcity limiting a class of models. The intervention can successfully transform the relation between theoretical complexity and measurement resolution. Yet the resulting abundance of high-dimensional data may make interpretation, storage, provenance, or validation newly central. The earlier contradiction has not simply been "solved" in a way that leaves the rest of the field unchanged. Its transformation produces a new state with a different contradiction structure.

The process can be represented schematically as

$$(S_t, \mathcal{C}_t) \xrightarrow{a_t} (S_{t+1}, \mathcal{C}_{t+1}),$$

where $\mathcal{C}_t = \mathcal{C}(S_t)$ and $\mathcal{C}_{t+1} = \mathcal{C}(S_{t+1})$. The important point is that $\mathcal{C}_{t+1}$ should be reconstructed rather than inferred by mechanically updating the previous list. Some earlier tensions may disappear, some may persist in altered form, some may change structural importance, and others may emerge only because the intervention created new capabilities or dependencies.

This requirement exposes a weakness in static planning. A long research plan often assumes that later tasks can be specified accurately before earlier transformations occur. Historical development limits such foresight. The completion of an intermediate stage can reveal phenomena, constraints, or opportunities that were not identifiable from the earlier state. A historically self-evolutionary AI therefore needs a capacity for planned incompleteness: it can define milestones and conditional branches while recognizing that later problem formulation must be reopened after material feedback is obtained.

DRAFT

Emergent contradictions also complicate evaluation of success. An intervention can be successful with respect to the contradiction it targeted while simultaneously creating a new limitation. This does not necessarily count as failure. The relevant question is whether the new state constitutes a defensible development of inquiry, whether the new tensions are tractable, and whether important capacities or safeguards were lost in the transition. Scientific development can advance by transforming the form of its difficulties rather than by reducing their number monotonically.

For AI systems, this implies that post-intervention analysis must be more than checking whether the immediate objective was achieved. The system should compare predicted and observed consequences, update its reconstruction of the state, identify newly relevant dependencies, and explicitly consider whether the evaluator used before the intervention remains adequate. The last issue is postponed to the later section on evolving methods and criteria, but its basis appears here: once intervention changes the problem state, the standards used to judge the previous state may no longer fully describe what matters.

7.6 Re-Problematization and Recursive Historical Inquiry

Dialectical transformation culminates not in closure but in re-problematization. After an intervention changes the research state, the problem itself must be posed again. This step is conceptually distinct from simply rerunning the same optimization procedure with new data. Re-problematization asks whether the original question retains the same meaning, scope, priority, and evidential requirements under the transformed conditions.

A long-standing question can survive linguistically while changing scientifically. "What causes this phenomenon?" can refer to different causal structures after measurement resolution improves. "Which model is better?" can change when a new dataset reveals that the models address different regimes. "How can this process be automated?" can become a question about oversight or exception handling after routine operations have already been automated. The persistence of words can conceal transformation in the underlying research object.

A historically self-evolutionary AI should therefore perform a recursive sequence of reconstruction, diagnosis, intervention, observation, and renewed reconstruction. Schematically,

$$S_t \to \mathcal{C}(S_t) \to A_t \to S_{t+1} \to \mathcal{C}(S_{t+1}) \to \cdots ,$$

where A_t denotes a set of candidate interventions rather than a predetermined unique action. Each cycle can alter both the state and the problem representation. The sequence is historical because later reasoning depends on the trajectory that generated the current state, not only on the current snapshot.

Re-problematization also creates a role for scientific judgment that cannot be reduced to autonomous optimization. Researchers may disagree about whether an intervention transformed the relevant contradiction, whether a newly visible tension is central, or whether the research programme should continue along the same trajectory. Such disagreement is not a defect to be eliminated automatically. It can provide evidence that the state reconstruction is incomplete or that multiple legitimate developmental paths remain open. An AI designed for historical inquiry should make these disagreements legible, preserve alternative reconstructions, and support comparison rather than prematurely collapse them into one authoritative narrative.

This recursive structure also limits the ambition of the present proposal. Historically self-evolutionary AI is not defined here as a machine capable of forecasting the entire development of science. Historical evolution is partially observable, contingent, institutionally distributed, and affected by events outside the system's model. The stronger and more defensible requirement is that the AI represent its problem-solving activity as participation in an evolving research state, recognize when its own interventions change that state, and reopen the problem accordingly.

The resulting account of dialectical transformation can now be stated compactly. A scientific contradiction is diagnosed within a historically reconstructed research state. Candidate interventions are compared according to the relations they alter, the elements they negate or preserve, the mediations they construct, and the future research possibilities they open or close. Intervention produces a new state

DRAFT

whose contradiction structure cannot be assumed to match the old one. The system therefore reconstructs the state and re-problematizes the inquiry. Scientific development becomes a recursive history of transformations rather than a sequence of answers to an invariant question.

This account prepares the next step of the argument. If interventions are valuable partly because they alter future reachability, then apparently modest research problems can possess developmental importance far beyond their immediate results. The next section examines such intermediate problems and develops the idea of the generativity of research.

8 Intermediate Problems and the Generativity of Research

Section 7 argued that scientific interventions should be assessed partly through the research states they help to produce. This claim changes the status of intermediate problems. In an answer-centered model of research, an intermediate problem is usually treated as a subordinate decomposition of a larger task: a difficult problem is divided into smaller pieces, the pieces are solved, and their solutions are recombined. Such decomposition is often useful, but it does not capture the historical role that modest research can play. An intermediate problem can matter because its solution changes the conditions under which later problems can be formulated, investigated, or established. It can create evidence, methods, instruments, standards, classifications, collaborations, or negative knowledge that did not previously exist. In that sense, the result is not only an answer to a smaller question. It is a transformation of the research state.

The term *generativity* is used here for this developmental property. A research action is generative when it contributes to the emergence of new capacities for inquiry, especially when those capacities enlarge, reorganize, or clarify the future research space. Generativity should not be confused with novelty, productivity, or the number of subsequent projects that happen to cite a result. A technically novel result may be developmentally sterile, while an apparently modest standard, dataset, calibration method, or conceptual distinction can reorganize a field. Nor does generativity imply that research should always maximize the number of future options. Some constraints are epistemically or ethically valuable, and some proliferations of possibility make validation harder rather than easier. The concept is therefore analytical before it is evaluative: it directs attention to how research changes the conditions of later research.

This perspective is particularly important for artificial intelligence intended to operate over long scientific horizons. A system optimized only for immediate task success may repeatedly select projects that produce locally strong outputs while neglecting the intermediate work needed to make more difficult questions tractable. Conversely, a system that merely generates increasingly ambitious hypotheses may produce what this paper calls epistemic stage-skipping: conclusions whose conceptual form can be produced before the evidential, methodological, or institutional chain required to establish them has been built. Historically self-evolutionary problem solving requires a different temporal orientation. It must reason about which smaller problems can productively transform the state from which harder problems will later be approached.

8.1 Intermediate Problems as Developmental Objects

An intermediate problem is often described instrumentally: problem p_1 matters because it is a component of a larger problem P. The relationship can be real, but it is not the only one. A research problem can also be intermediate because of its position in a historical sequence of capability formation. Its significance lies in what becomes possible after it is investigated.

Suppose a long-term question P is conceptually intelligible at time t but is not presently establishable. The obstacle may be missing evidence, insufficient resolution, an unstable classification, lack of comparative data, or the absence of a method capable of distinguishing relevant alternatives. A smaller problem p_t may be accessible under the current state S_t. If work on p_t produces a transformation

$$S_t \xrightarrow{p_t} S_{t+1},$$

then the value of p_t cannot be inferred only from the proposition established at the end of the study. The study may also have created a dataset, clarified a variable, revealed a hidden confound, stabilized a protocol, or demonstrated that a class of measurements is inadequate. These changes alter the relation between the field and the long-term question.

This makes intermediate status historical rather than purely logical. A problem can be logically unrelated to the final wording of P and still be developmentally necessary for reaching it. The construction of an instrument, for example, may not appear as a sub-proposition of the theory eventually tested with that instrument. The creation of a shared ontology may not answer any substantive empirical question by itself, yet it can make multi-site comparison possible. A replication failure may not advance the original positive hypothesis, but it can force a redefinition of the phenomenon and thereby change the trajectory of the programme. Intermediate research is therefore better represented as state-transforming activity than as a set of nested subquestions.

The same problem can also occupy different developmental roles at different moments. A dataset that is generative when a field lacks common evidence may become constraining after its categories are treated as natural and reused beyond their domain. A benchmark that initially enables comparison can later narrow inquiry around what the benchmark happens to measure. Historical reasoning therefore cannot permanently classify a particular object as generative. It must ask what function the object performs within the present research state and how that function changes over time.

For AI, this implies a shift from decomposition alone to sequencing. A capable system should not only ask which subproblems are easier. It should ask which tractable problems, if investigated now, are likely to alter the conditions under which later inquiry proceeds. The relevant relation is not simply

$$P = p_1 + p_2 + \cdots + p_n,$$

but a historical sequence in which the outcome of one stage changes the structure of the next. Later problems need not be fully specifiable in advance because their formulation can depend on what earlier stages reveal.

8.2 Generative Value Beyond Immediate Results

The developmental significance of intermediate research creates a distinction between immediate result value and generative value. Immediate value concerns what a project directly establishes, predicts, explains, measures, or builds. Generative value concerns the capacities for subsequent inquiry that the project helps to create or transform. These dimensions can align, but they need not.

Consider two hypothetical interventions. The first produces a small improvement on an established benchmark using a highly specialized method. The second produces no performance improvement but develops a reproducible measurement protocol that reveals previously hidden differences among benchmark cases. Under a narrow evaluator, the first intervention may appear superior. Historically, however, the second may change the field more profoundly because it reorganizes what later experiments can discriminate. Its contribution lies partly in the future research states it enables.

This distinction should not be reduced prematurely to a single numerical objective. One could imagine a quantity representing the increase in future reachability generated by an intervention, but such a quantity would immediately inherit difficult normative and epistemic questions. Which future questions count as valuable? How should one compare breadth of options with reliability, cost, or legitimacy? How far into the future should consequences be traced? How should irreversible lock-in be treated? The present argument therefore uses generative value as a comparative concept rather than a fully specified scalar.

A useful qualitative analysis asks several questions. Does the intervention create a reusable capacity rather than a one-off result? Does it reduce uncertainty about which later path is worth pursuing? Does it expose distinctions that were previously conflated? Does it improve the ability of independent groups to test claims? Does it preserve provenance and comparability across time? Does it make alternative explanations more discriminable? Does it create dependencies that narrow future inquiry? These questions describe different forms of generativity without assuming that they collapse into one metric.

DRAFT

Generative value also has a temporal structure. Some interventions have high short-term value and low long-term generativity; others have weak immediate payoff but strong delayed effects. Scientific institutions often struggle with this distinction because publication, funding, and evaluation cycles reward visible near-term outputs. A historically self-evolutionary AI should at least represent the mismatch. Otherwise it can reproduce the same short-horizon selection pressures at machine speed.

The point is not that long-term value should always dominate. A field sometimes needs decisive short-term results, and speculative infrastructure can consume resources without producing useful inquiry. The stronger claim is that an AI reasoning about scientific development should make the temporal profile of value explicit. It should be able to explain when a modest project is being recommended because of the state transformation it may enable rather than because of the immediate magnitude of its result.

8.3 Evidence, Methods, Instruments, and Infrastructure as Epistemic Production

The generativity of intermediate research becomes clearer when scientific outputs are understood more broadly than as propositions. Inquiry produces epistemic capacities in many forms. A new instrument changes what can be observed. A method changes what transformations of evidence can be performed. A dataset changes what comparisons can be made. A standard changes whether results from different sites can be aligned. A repository changes whether evidence can be accumulated and revisited. A governance procedure can change whether data can be shared legitimately across institutions. These products can be conditions of knowledge even when they are not themselves theoretical conclusions.

This observation connects directly to the historical-materialist orientation developed earlier. Productive capacities of inquiry are not merely background resources consumed by science. They are also outcomes of scientific activity. Research creates the means through which later research proceeds. Instruments, computational workflows, classification schemes, and institutional arrangements can therefore be understood as historically produced components of the research state.

For an AI system, this means that problem solving cannot be represented only as movement in a space of hypotheses. The system must sometimes recommend producing an epistemic object whose primary function is enabling later work. Examples include a calibration study before a large observational programme, a shared negative-control dataset before model comparison, an uncertainty taxonomy before automated decision support, or a protocol for documenting provenance before cross-institutional synthesis. Such projects can appear indirect if evaluation focuses only on the final substantive question. Under a developmental view, their indirectness is precisely their function.

Infrastructure deserves particular attention because it produces durable path dependence. Once a field adopts a dominant measurement platform, database schema, benchmark, or reporting standard, future research often reorganizes around it. This can be highly generative by lowering coordination costs and making cumulative work possible. It can also suppress alternatives by making nonconforming evidence expensive to collect or difficult to publish. A historically self-evolutionary AI should therefore treat infrastructure proposals as interventions with broad state-transforming effects rather than as neutral technical conveniences.

The same applies to conceptual infrastructure. Distinctions, taxonomies, and naming conventions can enable inquiry by making previously diffuse phenomena comparable. Yet conceptual stabilization also risks reification. Categories created for one research purpose can be mistaken for fixed properties of the world. Generative research should therefore preserve enough historical information to show why a category was introduced, what evidence it organizes, and under what conditions it should be reconsidered.

The general implication is that scientific productivity cannot be measured solely by the number of answers generated. A field develops by producing the means of its own further inquiry. AI that participates in long-horizon science must therefore reason about the production, maintenance, and revision of those means.

DRAFT

8.4 Negative Results and Constraint Discovery

Generativity is not limited to the creation of new capabilities. Research can also be generative by discovering constraints. A negative result, failed replication, impossibility proof, or demonstration of measurement insufficiency can reduce an apparently broad problem space in a scientifically productive way. Such work changes what should be attempted next.

This point matters because optimization-oriented systems can undervalue failure. If a research agent is rewarded primarily for producing positive findings or benchmark gains, unsuccessful experiments can appear as wasted iterations. Historically, however, a well-characterized failure can be a major state transition. It can eliminate a family of mechanisms, reveal that a variable is unidentifiable under available data, expose an instrumentation artifact, or show that two theories make no discriminable predictions under the present experimental regime.

Constraint discovery is especially important when a long-term problem is underdetermined. Suppose several theoretical frameworks can reproduce the same available observations. The immediate temptation is to generate increasingly elaborate arguments for choosing among them. A more generative intermediate problem may ask what new observation would actually discriminate the alternatives. If the answer is that no feasible observation under current instrumentation can do so, that negative conclusion is itself informative. It redirects effort from theoretical preference toward measurement development.

This kind of result transforms the reachable problem space by closing paths as well as opening them. The concept of generativity therefore should not be equated with expansion. Productive development can involve disciplined contraction: ruling out pseudo-problems, identifying impossible requirements, abandoning redundant distinctions, or recognizing that a claimed question cannot be answered at the current level of evidence. What matters is whether the transformation improves the structure of subsequent inquiry.

For historically self-evolutionary AI, negative results should therefore be stored with rich context rather than compressed into a binary record of failure. The system should retain what was attempted, under which assumptions, with which instruments, what kind of failure occurred, and which dependencies the failure revealed. Otherwise later generations of the system may repeatedly rediscover the same constraint or misinterpret an earlier failure as a universal impossibility when it was actually conditional on a historical state.

This requirement also links generativity to memory. Historical memory is valuable not merely because it preserves successful solutions but because it preserves the boundary conditions of unsuccessful ones. A developmental system needs both kinds of information to reason about what has become possible and what remains blocked.

8.5 Epistemic Stage-Skipping and Premature Closure

The concept of intermediate generativity helps clarify a specific risk of AI-assisted research: epistemic stage-skipping. Large language models can combine concepts drawn from many periods and disciplines within a single response. This capacity is valuable for analogy and synthesis, but it can produce an illusion of developmental readiness. A theoretical conclusion may be linguistically constructible even when the evidential and methodological sequence required to establish it has not occurred.

Stage-skipping should not be defined as the mere use of advanced concepts. Scientific progress often depends on importing mature methods from one field into another, and AI may accelerate such transfer productively. The problem arises when an output suppresses or bypasses the missing conditions that separate a plausible formulation from an establishable claim. A sophisticated theory can be proposed before the relevant variables are measurable, before the key distinctions are operationalized, before competing explanations can be discriminated, or before the data needed for validation exist.

In such cases, the appropriate response is not necessarily to reject the theory. The theory may be insightful. The historical question is what intermediate work would convert it from an attractive possibility into a researchable programme. A historically oriented AI should therefore be able to decompose epistemic distance into missing conditions. It might say that a proposed claim requires a measurement

DRAFT

method, a longitudinal dataset, a stable taxonomy, and a replication protocol before strong conclusions are warranted. The system would then search for tractable intermediate problems associated with those conditions.

Premature closure is the converse danger. An AI may generate a polished synthesis that appears to resolve a contradiction before the underlying research state has changed. Because language models are effective at producing balanced formulations, rhetorical synthesis can be mistaken for scientific transformation. The result can conceal unresolved empirical conflicts or material constraints behind conceptual integration.

This is why the dialectical account developed earlier insists on mediation and state change. A synthesis is developmentally meaningful only when it corresponds to a transformation in how the problem can be investigated, represented, or evaluated. If two competing theories are combined in prose but no new discriminating evidence, method, or conceptual structure results, the contradiction may remain exactly where it was.

An AI designed to avoid stage-skipping and premature closure would therefore attach developmental status to its outputs. It could distinguish exploratory synthesis, testable hypothesis, presently researchable claim, and presently establishable conclusion. These labels would not be universal epistemic ranks; they would be judgments relative to the reconstructed research state. Their purpose would be to make visible the gap between what can be said and what can responsibly be claimed.

8.6 Sequencing and Historical Research Strategy

Once intermediate problems are recognized as state-transforming, research strategy becomes a sequencing problem. The system must reason not only about what projects are valuable in isolation but about the order in which they should be attempted. Different sequences can produce different research states even when they contain similar component tasks.

A simple illustration is the difference between collecting a large dataset before stabilizing a measurement protocol and stabilizing the protocol before scaling data collection. The first sequence may produce more observations quickly but create long-term comparability problems. The second may delay immediate output while producing a more reliable infrastructure for cumulative evidence. Which sequence is preferable depends on cost, reversibility, uncertainty, and the role of the data in later inquiry.

Historical strategy therefore differs from static project ranking. A project with moderate standalone value can become highly valuable when performed before another project because it reduces uncertainty or prevents lock-in. Conversely, an apparently strong project can become wasteful if performed before a prerequisite distinction has been clarified. The order of interventions forms part of their epistemic meaning.

This also motivates conditional planning. Because future states are only partially predictable, a historically self-evolutionary AI should avoid specifying an entire long-term sequence as if each later step were already known. A better plan contains branching points and criteria for re-evaluation. After an intermediate result changes the state, the system reconstructs the relevant conditions and decides whether the next planned step remains appropriate.

Such planning can be represented without assuming deterministic control. Let A_t be a set of feasible interventions under S_t. The task is not to select a fixed path through all future actions, but to identify an intervention whose consequences are informative or generative enough to justify the next reconstruction. This places value on experiments that reduce uncertainty about the direction of the research programme, not only those that directly maximize performance.

For AI-assisted academia, this perspective suggests a different interaction pattern from one-shot question answering. When presented with an ambitious problem, the system can return a developmental map: what is presently known, which constraints appear structurally important, what intermediate studies would change those constraints, what evidence would trigger different next steps, and which parts of the plan should remain deliberately open. The output becomes a historically situated research strategy rather than a synthetic final answer.

DRAFT

8.7 Limits of Generativity as a Criterion

Generativity is useful precisely because it draws attention to consequences that immediate evaluation misses. It should not become an all-purpose criterion of scientific worth. Several limitations are important.

First, generativity can be epistemically ambiguous. A new method can generate many research programmes because it is genuinely powerful, or because it encourages repeated application to poorly chosen problems. A benchmark can organize a field while simultaneously narrowing imagination. Citation cascades and institutional adoption can therefore indicate social amplification rather than epistemic value.

Second, generativity can conflict with ethical and political constraints. A research intervention may greatly expand technical capabilities while creating unacceptable risks, concentrating control, violating rights, or undermining legitimate forms of oversight. The fact that an action increases future reachability does not make the increase desirable. Historical-material analysis should make these relations visible rather than smuggle them into a single optimization target.

Third, generative effects are difficult to forecast. Scientific development is contingent, and the value of an intermediate result can become apparent only retrospectively. An AI should therefore represent uncertainty around generative judgments and avoid presenting speculative future influence as a precise prediction. Multiple plausible developmental trajectories may need to be preserved.

Fourth, the pursuit of generativity can itself create lock-in. If an AI repeatedly selects projects that strengthen one research infrastructure because that infrastructure makes further work easier, it can produce a self-reinforcing path in which alternatives become progressively less accessible. A system should therefore consider option preservation and periodically ask whether the current trajectory is being followed because it remains epistemically justified or merely because previous investments made it convenient.

Finally, some research is valuable without being strongly generative. Descriptive scholarship, careful replication, preservation of historical evidence, resolution of a local uncertainty, or clarification of a bounded question can be worthwhile even if it does not open a large future research space. The present argument does not subordinate all scientific value to developmental leverage. It identifies generativity as one dimension that becomes especially important when AI is asked to plan and participate in long-horizon problem solving.

The resulting picture can be summarized as follows. Intermediate problems matter not only because they are easier fragments of larger problems but because they can transform the conditions under which later inquiry proceeds. Their outputs include evidence, methods, instruments, standards, infrastructures, constraints, and historical knowledge. Their value can therefore lie in the research states they make possible, the paths they rule out, and the uncertainty they remove from subsequent sequencing. This perspective explains why a self-evolutionary scientific AI should sometimes recommend a modest study instead of attempting the most ambitious available theory. The relevant question is not simply which problem appears most important in the abstract, but which tractable intervention can produce a defensible next state of inquiry.

With this account in place, the argument can turn from the developmental role of individual research steps to the architecture of the overall problem-solving process. The next section assembles historical state reconstruction, contradiction analysis, intervention comparison, generative sequencing, and re-problematization into a unified account of historically self-evolutionary scientific problem solving.

9 Historically Self-Evolutionary Problem Solving

The preceding sections have treated scientific inquiry as a historically conditioned process in which the research state, the structure of contradictions, and the reachability of future questions can all change. This section brings those elements together into a single philosophical model of historically self-evolutionary problem solving. The purpose is not to specify a software architecture or to claim that dialectical and historical-materialist concepts can be translated directly into executable procedures. It is to clarify the

DRAFT

sequence of judgments that an artificial system would need to approximate if it were to participate in long-horizon scientific development rather than merely optimize answers inside a supplied problem frame.

The central difference can be stated simply. A conventional problem-solving system receives a problem and searches for a solution. A historically self-evolutionary system must also reconstruct the conditions under which the problem has become formulable, determine which parts of the problem are presently reachable, identify the relations that constrain further development, choose interventions that can transform those relations, and then reconsider the problem after the intervention changes the state of inquiry. The unit of reasoning is therefore not a static pair consisting of a problem and a candidate answer. It is a temporally extended process in which problems, conditions, and criteria of progress may all be revised.

A schematic cycle is

$$S_t \to \mathcal{Q}(S_t) \to \mathcal{C}(S_t) \to A_t \to S_{t+1} \to \mathcal{Q}(S_{t+1}) \to \cdots,$$

where S_t denotes the reconstructed research state, $\mathcal{Q}(S_t)$ the questions or research moves presently reachable from that state, $\mathcal{C}(S_t)$ the relevant contradiction structure, and A_t a set of candidate developmental interventions. The arrows do not represent deterministic transitions. Each stage contains uncertainty, interpretation, and potentially several legitimate representations. Their role is to make explicit that historically self-evolutionary problem solving repeatedly returns from action to reconstruction rather than assuming that the initial problem formulation remains adequate throughout the process.

9.1 Historical Reconstruction as the Starting Operation

Historically self-evolutionary reasoning begins from reconstruction rather than immediate solution generation. When presented with a difficult scientific question, the system first needs to determine what kind of problem is actually being posed in the present historical state. This requires more than retrieving relevant publications. The system must reconstruct, at an appropriate level of abstraction, the concepts available to the field, the evidence already produced, the instruments and methods through which new evidence can be generated, the institutional arrangements under which work is coordinated, and the unresolved dependencies that shape what can be attempted next.

Section 5 represented this reconstruction schematically as $S_t = \langle X_t, \Gamma_t, H_t \rangle$. The usefulness of that representation lies in the distinction among relevant elements, their relational organization, and the historical trajectory through which the current configuration emerged. The same collection of resources can support different research possibilities when organized differently. A large dataset isolated behind incompatible governance arrangements is not equivalent to the same dataset embedded in a trusted and reproducible collaboration. A measurement technology without a shared calibration standard is not equivalent to the same device situated within a mature measurement practice. Historical reconstruction therefore needs to include relations of access, coordination, validation, and dependence rather than merely the existence of components.

The reconstruction should also be problem-relative. A complete representation of a scientific field is neither possible nor necessary. The relevant state is the one needed to understand why a given question is presently accessible, blocked, underdetermined, or prematurely formulated. A question about causal mechanism may require a different reconstruction from a question about measurement reliability even when both occur in the same field. The system should therefore treat the state representation as a revisable explanatory model rather than as an exhaustive ontology of science.

This starting operation has an epistemic consequence. Before asking how to solve the stated problem, the system may discover that the problem has been framed at the wrong developmental level. A request for a unified theory may rest on variables that are not yet measurable. A request for comparative evaluation may presuppose datasets that are not commensurable. A request for automated discovery may presuppose an evaluator whose validity has not been established. In such cases, historical reconstruction changes the immediate task. The first scientifically responsible output may be a diagnosis of missing conditions rather than a proposed answer.

DRAFT

This is one place where the distinction between linguistic availability and historical readiness becomes important. Large language models can often formulate advanced hypotheses because their training data compress traces of many historical stages into one accessible representational space. That capacity can support fruitful synthesis, but it does not guarantee that the present research state contains the evidence, instruments, or methods needed to establish those hypotheses. Historical reconstruction therefore acts as a discipline against epistemic stage-skipping: it asks what would have to exist in the research state for the attractive theoretical formulation to become a legitimate target of investigation.

9.2 Developmental Frontier and Present Reachability

Once a research state has been reconstructed, the next operation is to identify the developmental frontier. The frontier is not simply the most difficult unresolved question in a field. It is the boundary between what can presently be investigated with available capacities and what remains dependent on transformations that have not yet occurred.

This idea builds on the distinction in Section 4 among conceptual possibility, researchability, and establishability. A proposition may be conceptually expressible while remaining outside the present researchable set. A hypothesis may be researchable in the sense that relevant observations can be collected while remaining difficult to establish because the available evidence cannot discriminate among competing explanations. Historically self-evolutionary reasoning therefore needs a differentiated map of reachability rather than a binary label of solvable or unsolvable.

Let P denote a long-horizon scientific problem. The relevant question at time t is not only whether P can eventually be addressed, but which transformations separate the present state from a state in which P becomes researchable or establishable. The system may infer a sequence of conditional dependencies such as

$$S_t \xrightarrow{p_1} S_{t+1} \xrightarrow{p_2} S_{t+2} \xrightarrow{p_3} \cdots,$$

where the p_i are intermediate research problems, infrastructure-building tasks, conceptual clarifications, or empirical tests. The sequence is not a fixed plan generated once at the beginning. Each transition can reveal new information that changes the next step.

The developmental frontier is therefore better understood as a structured region of uncertainty. Some missing conditions are already identifiable: a calibration study, a shared ontology, a longitudinal dataset, or a method for discriminating two mechanisms. Other conditions are latent and become visible only after an intervention. A historically self-evolutionary system should represent both known dependencies and acknowledged unknowns. This prevents a research roadmap from being mistaken for a complete prediction of future development.

The frontier also has a social and institutional dimension. A field may possess the conceptual and technical capacity to address a problem while lacking an institutional arrangement that allows those capacities to interact. Relevant data can exist in several jurisdictions but remain unusable together. Expertise can be distributed across organizations whose incentives discourage collaboration. Evaluation can reward local benchmark gains while the developmental need is a shared infrastructure. In such cases, the frontier is not located only in theory or experiment. It lies in the relations through which scientific capacities are organized.

This broader conception of reachability changes what counts as an appropriate next problem. The system should not automatically choose the question closest in semantic content to the long-term goal. A seemingly indirect intervention can be more developmentally relevant if it changes several dependencies at once. The frontier is therefore defined by the structure of the research state, not by the linguistic distance between the current task and the final aspiration.

9.3 Structural Diagnosis and Developmental Priority

The reconstruction of the research state and its frontier provides the basis for contradiction analysis. Section 6 argued that a scientific contradiction, in the limited sense used here, is not merely a disagreement

DRAFT

or shortage. It is a relationally grounded tension that is reproduced within the organization of inquiry and has consequences for further development. Historically self-evolutionary problem solving needs to determine which such tensions deserve developmental attention at the present stage.

This does not require treating one contradiction as metaphysically fundamental. Structural centrality is temporal and problem-relative. A tension can be central because several other obstacles depend upon it, because it repeatedly regenerates failures across different projects, or because transforming it would change a large region of the reachable research space. Another tension can be scientifically serious but derivative in the sense that it is largely produced by the first. The distinction concerns organization, not moral worth.

A useful diagnosis therefore asks at least four kinds of questions. First, what processes reproduce the tension? Second, what other difficulties depend on or propagate through it? Third, which research capacities are prevented from interacting because of it? Fourth, how would the reachable problem space change if the tension were transformed? These questions move contradiction analysis away from rhetorical opposition and toward historical structure.

Consider a field in which model sophistication increases faster than the quality of available evidence. One response is to search for better models. Another is to improve data collection. Yet the deeper structural diagnosis may reveal that the field's evaluation system rewards increasingly complex models even when relevant measurements remain weak. In that case, the contradiction is not reducible to model complexity versus data scarcity. It includes an institutional relation that reproduces the mismatch. A developmental intervention aimed only at the model may therefore leave the source of the tension intact.

The same reasoning applies to apparently technical bottlenecks. Suppose multiple laboratories repeatedly fail to reproduce one another's measurements. The immediate problem might be described as statistical noise. Historical reconstruction may instead show that instruments are calibrated under incompatible procedures and that no institution maintains a shared reference standard. The structurally central contradiction then lies in the relation between distributed experimental capacity and absent standardization. A new statistical model can improve local estimates while leaving the developmental bottleneck untouched.

Developmental priority should therefore be inferred from relational consequences rather than from salience alone. Problems that attract the most attention are not necessarily the ones whose transformation will most alter the research state. Conversely, a low-visibility standard, method, or coordination mechanism can be highly central if many later investigations depend on it. A historically self-evolutionary AI should make this reasoning explicit so that users can contest both the reconstructed dependencies and the proposed priority.

DRAFT

9.4 Comparison of Developmental Interventions

After a structurally relevant contradiction or dependency has been identified, the system must compare possible interventions. This is the point at which historically self-evolutionary reasoning differs most sharply from selecting the candidate with the highest immediate score. The intervention is evaluated by the state transition it may produce, including both intended and unintended changes in future inquiry.

Let $a \in A_t$ denote a candidate intervention. The relevant object is not only the immediate result of a, but the transition

$$S_t \xrightarrow{a} S_{t+1}^{(a)}.$$

Different actions can produce similar local outcomes while generating very different subsequent states. One intervention may yield a strong result that depends on a proprietary dataset and an opaque evaluation procedure. Another may yield a weaker immediate result while producing a reusable dataset, transparent protocol, and reproducible benchmark. The first can dominate a short-horizon metric while the second has greater developmental significance.

The comparison therefore needs several dimensions that should not be collapsed prematurely into a single objective. One is epistemic effect: what uncertainty is reduced and what discriminations become possible? Another is capacity formation: what new tools, data, methods, or coordination mechanisms

are created? A third is dependency structure: does the intervention create a fragile bottleneck or distribute capability more broadly? A fourth is reversibility: if the intervention proves misguided, can the field recover alternative paths? A fifth is normative admissibility: are the resulting research possibilities compatible with ethical and legal constraints that should remain binding?

These dimensions also show why dialectical transformation cannot be equated with choosing a compromise between two sides. An intervention can preserve an important function of the existing state while reorganizing the relations that made the conflict recurrent. A privacy-preserving data infrastructure, for example, does not simply average between data access and confidentiality. It attempts to transform the relation through a mediating arrangement. Whether it succeeds depends on technical, institutional, legal, and epistemic details, but the form of reasoning is developmental rather than merely aggregative.

The system should also compare interventions against plausible alternatives, including inaction where appropriate. Historical development is not improved by change as such. Some contradictions reflect protections that should constrain scientific expansion. Some proposed infrastructures create lock-in that is worse than the bottleneck they solve. Some costly intermediate projects have speculative generative value but weak evidence that they will produce the expected state transition. A historically self-evolutionary system therefore needs to distinguish generativity from indiscriminate expansion.

This comparison is necessarily uncertain. The system cannot know the full consequences of an intervention before it occurs. Its role is to make the hypothesized transition explicit: which elements of S_t are expected to change, which relations are expected to be reorganized, which capacities should be preserved, and what new dependencies may emerge. Such explicitness turns intervention choice into a contestable scientific judgment rather than an opaque optimization output.

9.5 Intermediate Sequencing and Material Feedback

The selected intervention is rarely the end of the process. Long-horizon research requires sequences of intermediate steps whose ordering matters because each step changes the conditions under which later steps are interpreted. Section 8 developed this point through the generative value of intermediate problems. Historically self-evolutionary problem solving integrates that idea into an iterative sequencing process.

A sequence should be understood conditionally rather than as a rigid plan. The system may formulate a local roadmap such as: first establish measurement reliability, then test whether two candidate mechanisms are empirically discriminable, then decide whether a larger intervention is justified. Yet each stage includes a branching condition. If measurement reliability fails, the next problem becomes instrumentation or protocol design. If the mechanisms remain observationally equivalent, the next task becomes discrimination rather than theory selection. If the expected distinction disappears under improved data, the original long-horizon problem may need reformulation.

Material feedback is what prevents this sequence from becoming purely textual self-reflection. A historically self-evolutionary AI must be capable, directly or through human and institutional collaborators, of incorporating observations that can contradict its prior reconstruction. Experimental outcomes, failed replications, cost overruns, inaccessible data, governance refusals, and unexpected side effects are all part of historical development. They can reveal that the assumed state transition did not occur or that a different contradiction has become central.

This point is especially important for self-improving AI. An agent that evaluates its own proposals primarily through internally generated critique can become increasingly coherent without becoming more historically adequate. The system may refine a mistaken representation of the research state rather than expose it to evidence capable of overturning it. Historically self-evolutionary problem solving therefore requires feedback channels whose epistemic authority is not reducible to the agent's own prior model.

Intermediate sequencing also creates a role for stopping conditions. A generative project can continue producing ever more infrastructure or methodological refinement without approaching the scientific question that originally motivated it. The system should therefore periodically ask whether the sequence remains developmentally connected to the long-horizon problem, whether new evidence has changed the

DRAFT

value of the original goal, and whether resources should be redirected. Historical self-evolution includes the possibility of abandoning a trajectory rather than treating persistence as progress.

The sequence is thus recursive but not circular. Each iteration should produce a changed informational or material basis for the next judgment. If repeated cycles merely regenerate the same recommendation, the system should diagnose whether the contradiction has been misidentified, whether the proposed intervention lacks causal leverage, or whether the relevant transformation lies outside the system's present action space.

9.6 Re-Problematization after State Transition

The defining operation of historical self-evolution occurs after intervention. Once the research state changes, the system must reconstruct the problem rather than simply continue the original plan. Section 7 described this as re-problematization. Here it becomes the mechanism through which problem solving acquires historical depth.

Suppose an intervention succeeds in producing a new measurement capability. The initial question may have been whether two theoretical explanations can be distinguished. The new instrument can make that comparison possible, but it can also reveal a previously invisible third phenomenon. The appropriate research problem is now different from the one specified before the intervention. Continuing to optimize the original two-model comparison would waste the developmental information created by the new state.

Re-problematization therefore requires rebuilding at least three objects: the research-state representation, the contradiction structure, and the reachable question-space. The system should ask which assumptions in the previous model remain valid, which dependencies have disappeared, which new dependencies have emerged, and which formerly speculative questions have become testable. It should also record which aspects of the earlier state were transformed by design and which changed unexpectedly.

This operation gives historical memory a functional role. The previous representation should not simply be overwritten. Without a record of how S_t became S_{t+1}, the system loses the ability to distinguish persistent structure from contingent outcome. It also becomes harder to evaluate whether an intervention worked for the reasons originally proposed. A later section will examine historical memory and alternative developmental paths in greater detail; the present point is that re-problematization depends on preserving enough of the transition history to make the new state intelligible.

Re-problematization also protects against premature convergence. A high-performing solution under the old evaluator can cease to be relevant after the research state changes. A benchmark can lose discriminative power. A variable can become measurable in a new way. A theoretical distinction can prove unnecessary. A historically self-evolutionary AI must therefore be willing to revise not only its answer but the question to which the answer was supposed to respond.

The process can be summarized as

$$(S_t, P_t, \mathcal{C}_t) \xrightarrow{a_t, o_t} (S_{t+1}, P_{t+1}, \mathcal{C}_{t+1}),$$

where P_t denotes the problem representation and o_t the observations generated during or after intervention. The notation emphasizes that P_{t+1} need not be a more detailed version of P_t. It can be a restructured problem produced by the historical transition itself.

9.7 Three Layers of Self-Evolution

The resulting account suggests that the phrase "self-evolutionary AI" is ambiguous unless the evolving object is specified. At least three layers can be distinguished.

The first is *agent evolution*. The AI changes its own prompts, tools, policies, code, models, or internal strategies. Contemporary self-improving and evolutionary systems already explore parts of this layer, as reviewed in Section 3. Such evolution can make the agent more capable, but it does not by itself guarantee that the agent represents scientific development historically.

The second is *problem-representation evolution*. The system revises how it defines the scientific problem, which variables and distinctions it considers relevant, which evidence it treats as discriminative,

DRAFT

and which subproblems it regards as reachable. This layer is essential because scientific development frequently changes the problem before it changes the answer.

The third is *problem-condition evolution*. The system participates in, recommends, or coordinates interventions that change the material and relational conditions under which inquiry proceeds: data resources, instruments, experimental arrangements, standards, methods, institutions, interfaces, or relations of collaboration. At this layer, problem solving acts on the world that makes problems solvable.

These layers interact but should not be conflated. An agent can improve itself while operating inside an unchanged research regime. A problem representation can evolve through conceptual reflection while material conditions remain fixed. A field can transform its infrastructure while an AI's internal strategy changes very little. Historically self-evolutionary scientific intelligence, in the stronger sense developed here, concerns the coupling among all three layers.

The coupling can be represented schematically as

$$A_t \leftrightarrow P_t \leftrightarrow S_t,$$

where A_t denotes the state of the agent, P_t the current problem representation, and S_t the historically constituted research state. Development at one layer can induce revision at the others. A new experimental result changes S_t and may require a new P_t. A revised problem representation can reveal that the agent lacks a tool and therefore motivate change in A_t. An improved agent can identify a previously unnoticed institutional bottleneck and propose an intervention that changes S_t.

This three-layer view also clarifies what the present paper does not claim. Historical self-evolution is not equivalent to unconstrained autonomous action. An AI need not control laboratories, institutions, or research communities in order to reason historically about their transformation. Human authorization, legal governance, scientific accountability, and domain expertise remain essential wherever interventions affect real institutions or people. The philosophical claim is about the structure of adequate problem solving: if scientific solvability depends on historically changing conditions, then an AI designed for long-horizon research must represent those conditions and their possible transformation even when humans retain decision authority.

Nor does the model imply that every scientific problem requires large-scale historical reconstruction. Many tasks are appropriately solved by local optimization, direct calculation, or ordinary experimentation. The historical mode becomes most relevant when the target problem persists across research stages, when current methods repeatedly reproduce the same bottleneck, when the evaluator is unstable or contested, or when important progress depends on creating new capacities rather than selecting among existing answers.

The transition from conventional self-improvement to historically self-evolutionary problem solving is therefore not a matter of adding one more optimization loop. It is a shift in the object of intelligence. The system must reason about how a scientific problem is generated, constrained, transformed, and reconstituted through changes in the research world itself. The next sections develop two consequences of this shift that have so far remained only implicit: the historical evolution of methods and evaluative criteria, and the role of memory, branching, and alternative trajectories in preserving the intelligibility of long-horizon development.

10 The Evolution of Problems, Methods, and Evaluation

Historically self-evolutionary problem solving cannot be reduced to the repeated application of a fixed method to an evolving sequence of problems. Once the research state changes, the means by which a problem is investigated can also change, and the criteria by which progress is judged can become inadequate to the transformed situation. This creates a second-order difficulty. An AI that is capable of revising hypotheses but not methods remains dependent on a historically fixed repertoire of inquiry. An AI that can revise methods but treats evaluative criteria as permanently external to development can still mistake success under an inherited benchmark for progress in the transformed research state.

DRAFT

The stronger historical view developed in this paper therefore requires co-evolution among problems, methods, evidence, and evaluation.

This claim does not imply that scientific standards are arbitrary or that an agent should rewrite its objectives whenever results are inconvenient. On the contrary, evaluative revision is more demanding than ordinary optimization because it concerns the conditions under which optimization is meaningful. A benchmark, loss function, evidential threshold, or methodological convention can be revised responsibly only when the system can show why the existing criterion has become developmentally misaligned, what is preserved from it, what new information motivates the revision, and how alternative criteria would change the reachable paths of inquiry. Historical self-evolution thus requires a distinction between adaptation within an evaluative regime and justified transformation of the regime itself.

The historical-materialist significance of this claim is specific. Methods and evaluative regimes belong to the organization of scientific production: they mediate what observations can be generated, which distinctions can become operational, which results enter comparison, and which activities are recognized as progress. Their historical location matters because an arrangement that initially coordinates inquiry can later become a bottleneck. The relevant question is therefore not whether an inherited method or evaluator is universally valid or invalid, but whether its domain of adequacy still corresponds to the developed research state.

10.1 Problems as Historically Moving Targets

Scientific problems are often treated as stable objects whose solutions improve over time. This representation is appropriate for some tasks: a mathematical conjecture can remain formally stable while proof techniques improve, and an engineering target can remain fixed while candidate designs become more efficient. Yet many long-horizon scientific problems change as inquiry develops. New measurements alter which distinctions matter, new theories reorganize the relevant variables, and new infrastructures make previously invisible phenomena observable. The problem that exists after such a transition is not always the same problem with more information attached to it.

The historical state representation introduced in Section 5 makes this explicit. Let P_t denote the current representation of a target problem under research state S_t. After intervention and observation, the system moves to a new state S_{t+1} and may require a revised problem representation P_{t+1}. The relation between the two need not be one of simple refinement. A distinction central to P_t can disappear, a hidden variable can become central, or the original target can divide into several questions whose developmental roles differ.

This movement is familiar in the history of science. Historical philosophy of science has long emphasized that scientific development can alter not only accepted answers but the significance of problems, exemplary forms of explanation, and standards of successful inquiry [18, 3]. The present argument does not require a strong paradigm-relativist thesis. It requires only the weaker claim that a problem representation is partly constituted by the methods and evidential relations available at a particular historical stage.

For AI, this creates a risk of representational inertia. A system can continue to optimize the problem statement supplied at the beginning of a project even after accumulated evidence has made that statement obsolete. The more capable the optimizer, the more efficiently it can pursue the wrong object. Historically self-evolutionary intelligence therefore needs explicit mechanisms of re-problematization: periodic opportunities to ask whether the current problem still organizes the relevant evidence and whether a different representation would better capture the transformed state.

Re-problematization should not be triggered by mere dissatisfaction with performance. A problem is not historically obsolete simply because it is difficult. The stronger signal is structural: changes in S_t alter the relations that originally made P_t informative. A measurement becomes possible that the old formulation treated as latent; an assumed boundary between two mechanisms dissolves; a new institutional constraint makes the original experiment practically irrelevant; or a negative result eliminates an entire family of interpretations. These are reasons to revisit the problem representation because the historical basis of the problem has changed.

DRAFT

10.2 Methods as Historically Developed Capacities

Methods are often represented computationally as tools selected from a repertoire. Under this view, the scientific agent faces a problem, chooses among statistical tests, simulation techniques, experimental protocols, search procedures, or software tools, and then evaluates the result. Historically, however, methods are themselves achievements of inquiry. They are invented, standardized, criticized, institutionalized, automated, and sometimes abandoned. A research system's methodological repertoire is therefore part of the historically constituted state rather than an external toolbox.

This distinction matters because a method can become the object of development. A field may possess a scientifically important question but lack a measurement protocol capable of operationalizing the relevant variable. It may have adequate observations but lack a method for integrating heterogeneous datasets. It may have a useful method whose assumptions are systematically violated by the new research setting. In such cases, applying existing methods more efficiently cannot resolve the problem. The research action must create, modify, validate, or combine methods.

The historical-material orientation developed in Sections 4 and 5 is especially relevant here. A method does not consist only of an abstract procedure. Its practical use can depend on instruments, trained personnel, software, calibration practices, shared standards, access to data, and institutional routines. A new method therefore changes the research state only when these supporting relations become sufficiently organized for the method to function as a reproducible capacity. Method development is consequently both conceptual and material.

Historical development can also reverse the developmental role of a method. A standardizing procedure can initially enable cumulative comparison and later conceal variation that becomes scientifically important. A statistical or experimental technique can enlarge the reachable problem-space under one data-generating regime and later become a bottleneck when scale, instrumentation, or the relevant object changes. Such a transition does not require the conclusion that the earlier method was simply wrong. It may have been historically adequate for an earlier state while becoming insufficient for a later one. For self-evolutionary AI, diagnosing this transition is as important as inventing a new method, because continued efficiency within the inherited method can otherwise reproduce the very limitation that further development has exposed.

This perspective also changes the role of AI-generated methods. A language model can propose a new analytic procedure or experimental design in text long before the surrounding field possesses the resources to validate it. The proposal is then conceptually available but not yet an operational scientific method. Historically self-evolutionary AI should distinguish between methodological imagination and methodological establishment. The latter requires evidence that the method works under specified conditions, that its failure modes are understood, and that it can be reproduced or critically inspected by relevant communities.

Method evolution therefore has its own intermediate problems. Before a method becomes a general scientific capacity, inquiry may need to solve problems of calibration, robustness, transfer, interpretability, interoperability, or cost. These problems can have high generative value because resolving them enlarges the future space of research actions. The method itself becomes an epistemic infrastructure through which later questions become reachable.

10.3 Fixed Objective Functions and Historical Misalignment

Evolutionary AI is most straightforward when candidate quality can be evaluated by a stable objective. Program-search systems such as AlphaEvolve illustrate the power of this arrangement when outputs can be automatically tested and quantitatively scored [7]. Similar advantages arise whenever scientific agents operate against benchmarks, formal proof checkers, simulation targets, or well-defined experimental readouts. A stable evaluator converts exploration into a tractable selection process.

The difficulty considered here appears when the evaluator is historically embedded in the research regime that is itself changing. Let F_t denote the evaluative criterion used at stage t. Ordinary optimization assumes that candidate actions or solutions can be compared under F_t while the criterion remains

DRAFT

sufficiently stable. Historically self-evolutionary problem solving must allow the possibility that an intervention changes the conditions under which F_t was meaningful. In that case, continuing to optimize the same criterion can reward behavior that no longer corresponds to scientific progress.

A benchmark can lose discriminative power because performance saturates. A proxy variable can cease to track the phenomenon it was intended to represent after the measurement regime changes. A replication criterion designed for one experimental organization can become inadequate when research shifts to a different scale or data-generating process. A ranking system can incentivize local improvement while suppressing method-building work whose value is primarily generative. None of these cases shows that evaluation should be abandoned. They show that evaluators have domains of adequacy.

The historical problem is therefore not simply evaluator drift. It is the possible divergence between the evaluator and the transformed research state. A fixed criterion can become part of a contradiction structure when it continues to organize incentives, selection, or resource allocation after the capacities and problems it was designed to govern have changed. In such cases, improving performance under the criterion may reproduce the bottleneck rather than overcome it.

This yields a stronger conclusion than the familiar warning that fixed objectives can be brittle. If F_t has become historically obsolete, more effective optimization under F_t can deepen epistemic error. The system may become increasingly successful according to the inherited measure while moving farther from the transformed scientific aim. The relevant failure is historical misalignment between evaluator and research state. Historically self-evolutionary intelligence therefore requires a capacity to detect such misalignment before increased optimization pressure converts an inherited standard into a stronger mechanism of lock-in.

Historically self-evolutionary AI should therefore represent an evaluator not merely as a function but as a historically situated relation among a target, an observable proxy, a method of measurement, and an institutional use. This makes revision more disciplined. The system asks which relation has failed: whether the observable no longer tracks the target, whether measurement error has changed, whether the criterion is being used outside its validated domain, or whether the scientific aim itself has been reformulated.

10.4 Evidence Standards under Changing Research Conditions

Evaluation is not limited to numerical objectives. Scientific inquiry also depends on standards governing what counts as adequate evidence. These standards vary across domains and can change when measurement capacities, experimental structures, or inferential methods change. Social epistemology further emphasizes that evidential assessment is distributed across communities, institutions, and relations of expertise rather than performed by isolated reasoners [14]. A historically self-evolutionary AI therefore cannot model evidence quality as a timeless scalar detached from the practices through which evidence is produced and criticized.

Consider a field in which evidence initially consists of small observational datasets. A later historical state may include high-frequency measurements, automated experiments, richer metadata, or standardized multi-site protocols. The inferential possibilities have changed, but so have the relevant risks. Large datasets can reduce sampling uncertainty while increasing dependence on hidden preprocessing pipelines, measurement artifacts, or institutional selection effects. The standard of adequate evidence should therefore not be represented as monotonically increasing with data volume.

Similarly, a new instrument can increase sensitivity while introducing a new calibration problem. A more complex model can fit previously unexplained variation while making parameter identification harder. A multi-agent scientific system can generate broader critique while creating new dependencies on model correlation and shared training data. Each historical gain creates its own evidential obligations. The relevant question is not whether the new state contains "more evidence" but whether the relations among observation, inference, and claim remain defensible.

This point connects directly to contradiction analysis. Evidential standards can become structurally central when several research problems depend on them. If multiple hypotheses cannot be discriminated because the measurement model is unstable, improving theory generation may have little developmental

DRAFT

value. The historically central intervention may instead be to stabilize the evidential relation. Conversely, once that relation is stabilized, the evidential bottleneck can recede and another contradiction can become central.

For AI, this implies that evidence evaluation must be state-sensitive. A claim judged weak at t can become well supported at $t + 1$ because new measurements or methods change the evidential structure. The reverse is also possible: a result that once appeared robust can become questionable after a newly discovered confound or broader dataset reveals the limits of earlier inference. Historical self-evolution therefore includes retrospective revision. The system must be able to revisit past judgments when the conditions of evidential assessment change.

10.5 Criterion Revision and Preservation

Allowing evaluative criteria to evolve creates an obvious danger: a self-improving system could protect its preferred trajectory by weakening the standards that expose failure. A historically materialist orientation does not remove this danger. Indeed, because it treats evaluative standards as historically constituted, it must state more clearly what distinguishes responsible revision from opportunistic self-exemption.

The Hegelian notion of sublation offers a useful constraint at the conceptual level. Revision should specify what is negated, what is preserved, and why the new organization can account for the developmental limitation of the old one [17]. Applied to evaluation, this means that a revised criterion should not simply discard inconvenient evidence. It should preserve the valid domain of the previous criterion where that domain remains intact, explain the failure that motivates revision, and make the transition inspectable.

A historically self-evolutionary AI should therefore retain evaluative lineage. If F_t is replaced or supplemented by F_{t+1}, the system should preserve the earlier criterion, the evidence supporting the change, the cases on which the two criteria disagree, and the expected consequences of adopting the new one. This turns criterion evolution into a comparative historical judgment rather than silent objective mutation.

The same requirement applies to methodological lineage. If M_t is replaced, bounded, or incorporated into M_{t+1}, the system should retain the conditions under which M_t was established, its validated domain, the supporting instruments and practices on which it depended, the limitation that motivates revision, and the relation between results produced under the old and new methods. Methodological evolution should therefore remain historically reconstructable. Otherwise a field can gain a new procedure while losing the ability to understand why earlier evidence was produced differently or where comparability has been broken.

The transition can be represented schematically as

$$(S_t, P_t, M_t, F_t) \xrightarrow{a_t, o_t} (S_{t+1}, P_{t+1}, M_{t+1}, F_{t+1}),$$

where M_t denotes the operative methodological repertoire and F_t the evaluative regime. The notation does not imply that all four components must change at every step. Its purpose is to show that a major research transition can alter several of them together.

Criterion revision also requires plural comparison. A new evaluator should, where feasible, be assessed against preserved alternatives rather than immediately becoming the sole basis of selection. The AI can ask whether a proposed criterion improves discrimination, whether it creates new blind spots, whether conclusions remain robust under multiple reasonable evaluators, and whether some disagreement reflects a genuine normative or scientific choice that should remain visible rather than collapsed into a single score.

This is particularly important when evaluation includes social or institutional objectives. A criterion can encode priorities about cost, access, safety, fairness, openness, or acceptable risk. These are not merely technical parameters to be optimized away. Historically self-evolutionary AI must distinguish scientific revision from changes that require human normative authorization. The fact that a criterion constrains research does not by itself make the criterion a contradiction to be overcome.

DRAFT

10.6 Co-Evolution of Problem, Method, and Evaluator

The preceding analysis suggests that long-horizon scientific development is better represented as a coupled process than as optimization over a fixed task. Problems define what is being investigated, methods determine what transformations and observations are available, and evaluators determine how outcomes are compared. Yet each component can alter the others.

A new method can reveal that the old problem combined two phenomena that should be separated. A reformulated problem can make an established benchmark irrelevant. A revised evidential standard can expose the inadequacy of a previously successful method. A new evaluator can incentivize infrastructure-building work that changes the research state. The developmental unit is therefore not any one component but the relational configuration among them.

This co-evolution can be understood through the research-state framework. At time t, the system reconstructs S_t, together with a problem representation P_t, methodological repertoire M_t, and evaluative regime F_t. It diagnoses contradictions and proposes interventions. After intervention, new observations o_t produce a transition. The next step is not simply to rerun the optimizer under the old configuration. It is to test whether P_t, M_t, and F_t remain adequate to S_{t+1}.

The resulting cycle can be summarized as

$$(S_t, P_t, M_t, F_t) \to \mathcal{C}_t \to A_t \to (S_{t+1}, P_{t+1}, M_{t+1}, F_{t+1}) \to \cdots .$$

The expression is intentionally schematic. It does not prescribe a computational architecture or imply that historical development follows a deterministic state-transition law. It identifies the objects that a future system would need to keep revisable if it is to participate in scientific development rather than merely search within a frozen representation of it.

This co-evolutionary view also clarifies why historical self-evolution should not be identified with unlimited self-modification. Some components should change slowly because their stability supports cumulative inquiry. Shared measurement standards, safety constraints, provenance rules, and reproducibility requirements can be valuable precisely because they resist local optimization pressure. Historical intelligence therefore includes judgment about what should remain stable long enough for learning to accumulate.

The relevant capacity is not maximal plasticity but differentiated revisability. The AI should know which parts of its problem-solving regime are currently hypotheses, which are validated methods, which are institutional conventions, which are normative constraints, and which are provisional evaluators. Different categories require different burdens of justification for revision. A system that changes everything continuously loses historical continuity; a system that changes nothing cannot adapt to historical development.

Historically self-evolutionary scientific intelligence therefore requires a balance between preservation and transformation. Problems, methods, and evaluative criteria must remain revisable, but their revision must be historically documented, empirically motivated, and open to contestation. The next section turns to the memory structure required for this balance. If development is branching and criteria can change, the system must preserve enough of its prior trajectories to reconstruct why a present state exists, which alternatives were abandoned, and whether forgotten paths should become reachable again.

11 Historical Memory and Alternative Developmental Paths

Historically self-evolutionary problem solving requires more than the capacity to move from one research state to another. It also requires a memory structure capable of preserving why a transition occurred, which alternatives were considered, which constraints were discovered, and how the criteria of inquiry changed along the way. Without such memory, the recursive cycle developed in Sections 9 and 10 becomes historically shallow. The system may adapt to a new state while losing the relations that make the new state intelligible. It can then repeat discarded mistakes, mistake contingent conventions for necessities, or interpret the current configuration as though it had emerged without alternatives.

DRAFT

The relevant memory is therefore not an archive in the ordinary sense of storing all previous outputs. Historical memory is a structured account of developmental lineage. It connects states, interventions, observations, problem reformulations, methodological changes, and evaluative revisions. It also preserves selected unrealized alternatives: paths that were rejected, deferred, blocked by material conditions, or rendered unattractive by criteria that may themselves later change. Such alternatives matter because historical development is branching rather than merely cumulative. A path that was irrational under S_t can become relevant under S_{t+k} when instruments, evidence, resources, or institutional relations change.

This section develops six implications of that claim. First, historical memory must preserve relations of derivation rather than only retained conclusions. Second, failed paths and negative results should be stored as situated constraints rather than erased as unsuccessful search. Third, alternative trajectories should remain representable even when one branch becomes dominant. Fourth, path dependence and lock-in must be treated as properties of the evolving research state. Fifth, counterfactual reconstruction can help determine whether an abandoned path has become reachable under changed conditions. Finally, historical memory requires selective retention: an intelligence that records everything without structure can become no more historically capable than one that remembers too little.

11.1 Historical Memory as Structured Lineage

A conventional memory system can store propositions, documents, observations, or prior interactions. A historically structured memory must additionally represent relations among them. It should be able to recover not only that a current claim is accepted, but which evidence supported it, which alternative interpretation it displaced, which method generated the evidence, and which assumptions defined the problem at that stage. Historical intelligence therefore depends on provenance across transformations.

Let H_t denote the developmental history associated with research state S_t. In the minimal representation introduced in Section 5, H_t was included to prevent the present research state from being treated as self-explanatory. We can now make its role more explicit. H_t should contain selected transition records of the form

$$\langle S_i, P_i, M_i, F_i, a_i, o_i, S_{i+1} \rangle,$$

where P_i is the operative problem representation, M_i the methodological repertoire, F_i the evaluative regime, a_i an intervention, and o_i the resulting observation or feedback. The notation is schematic. Its purpose is to emphasize that historical memory should preserve the relation between a decision and the conditions under which that decision was reasonable.

This differs from merely saving successive snapshots. Suppose a system records that method M_1 was replaced by M_2. Without the transition record, a later agent may infer that M_2 is simply superior. Yet the replacement may have occurred because a particular instrument made M_2 practical, because M_1 failed on a newly important class of data, or because a funding structure favored one experimental design. If those surrounding conditions later change, the historical meaning of the transition changes as well. A method that was reasonably abandoned can become useful again in a different state.

Structured lineage also supports contestability. Scientific development is rarely represented by a single uncontested sequence of justified decisions. Different communities can interpret the same transition differently, especially when the transition involved uncertain evidence, institutional authority, or competing evaluative criteria. Historical memory should therefore preserve disagreement where it is developmentally relevant. A later system should be able to distinguish between "this alternative was empirically excluded" and "this alternative lost institutional support under the criteria then in use." Those are different historical relations and should not collapse into the same status of rejection.

The need for lineage is consistent with the broader philosophy-of-science literature reviewed earlier. Accounts of scientific change associated with Kuhn emphasize that changes in exemplary problems, standards, and disciplinary organization can alter the meaning of later judgments about earlier science [3, 18]. The present argument does not require treating successive research states as mutually incommensurable. It requires only that later evaluation preserve enough context to avoid reading the past entirely through the categories of the present.

DRAFT

For AI, this creates a practical conceptual requirement. A memory useful for historical self-evolution should answer at least four kinds of query: how the current state arose; why a previous path was abandoned; which assumptions were necessary for a transition; and which alternatives remain conditionally reopenable. These are not retrieval questions alone. They are reconstruction questions about relations across time.

11.2 Failed Paths as Situated Constraints

Evolutionary and search-based systems often distinguish successful candidates from unsuccessful ones through selection. Open-ended systems complicate this picture by preserving diverse stepping stones rather than only the current best candidate. The Darwin Gödel Machine, for example, maintains an archive of prior agents because an apparently inferior agent can later serve as a productive stepping stone for a different trajectory [26]. POET similarly demonstrates that progress in open-ended search can depend on transfers among environments and solutions whose value is not visible under a single linear objective [20]. These systems already show why historical self-evolution should not identify memory with preservation of winners.

Scientific inquiry provides an even stronger reason to retain failed paths. A failed experiment, unsuccessful method, or rejected hypothesis can change what is rational to attempt next. Section 8 treated negative results as potentially generative because they can close unproductive regions of inquiry or reveal hidden constraints. Historical memory is what allows that generativity to persist beyond the local episode in which the failure occurred.

The key qualification is that failure must be stored as situated rather than universal. If intervention a_i failed under state S_i, the warranted historical record is not simply

$$a_i = \text{failure}.$$

The stronger record is closer to

$$(a_i, S_i) \mapsto o_i,$$

where o_i includes the observed mode of failure and the conditions under which it occurred. The difference matters because a later research state can remove the constraint responsible for the earlier failure. A technique that was too expensive can become cheap; a measurement that lacked resolution can become feasible; a hypothesis that could not be discriminated can become testable after a new instrument appears.

This state-sensitive treatment of failure prevents two opposite errors. The first is repetitive rediscovery: without historical memory, the system repeatedly attempts paths whose failure conditions remain unchanged. The second is permanent foreclosure: with an overly rigid memory, the system treats all previous failures as timeless prohibitions. Historically self-evolutionary reasoning needs a middle position. It should preserve why a path failed and periodically test whether the relevant failure conditions still obtain.

Negative results also have institutional histories. Some failures are widely published and integrated into disciplinary memory; others remain local to laboratories, unpublished repositories, abandoned codebases, or personal expertise. The social distribution of memory therefore affects the future reachability of research. A field can appear to have "forgotten" a path not because nobody attempted it, but because the evidential trace was not preserved in a form accessible to later inquiry. The social dimensions of scientific knowledge are relevant here: what becomes collectively available depends on communicative practices and institutional arrangements, not only on individual cognition [14].

A historical AI should consequently distinguish absence of evidence from evidence of historical non-attempt. It should avoid inferring that an unrepresented path was never explored. This is especially important when archives are uneven, proprietary, geographically concentrated, or biased toward positive outcomes. Historical reconstruction remains partial and fallible for the same reason that the research state itself is only partially observable.

DRAFT

11.3 Branching Trajectories and Preserved Alternatives

A linear historical representation has the form

$$S_0 \to S_1 \to S_2 \to \cdots \to S_t.$$

This is useful for displaying the realized sequence of dominant states, but it conceals the alternatives available at each transition. A more adequate representation is a branching structure in which a state can support several plausible interventions and therefore several possible successor states. Only one or a few branches may be realized, while others remain hypothetical, deferred, or institutionally excluded.

For historical self-evolution, the distinction between realized and unrealized branches is important. An unrealized branch is not equivalent to a failed branch. It may never have been attempted because resources were unavailable, because a competing programme attracted attention, because institutional incentives favored a different method, or because the problem was framed in a way that made the alternative invisible. A later research state can alter those conditions.

The system should therefore preserve a limited set of alternative developmental records. These records need not contain complete simulated histories. They can encode the intervention considered, the reasons for rejection or deferral, the anticipated requirements, and the uncertainties that prevented selection. In this sense, historical memory contains not only "what happened" but selected information about "what almost happened" and "what could not yet happen."

This does not make the model counterfactually omniscient. Alternative histories cannot be reconstructed as if their outcomes were known. A counterfactual branch should remain marked by its epistemic status. Some branches are strongly constrained by known mechanisms; others are speculative. Preserving this distinction is necessary to avoid converting imaginative alternatives into fabricated historical facts.

Branching memory also provides a safeguard against premature convergence. If a research agent repeatedly selects the same kind of intervention, the surrounding research state can become increasingly adapted to that intervention. Tools, benchmarks, datasets, and expertise accumulate around the selected branch. Later comparisons then favor the branch partly because previous selections made it easier to pursue. The resulting dominance may reflect genuine scientific success, historical reinforcement, or both. Preserved alternatives make this process visible.

The analogy with open-ended evolutionary AI is useful but limited. Systems such as POET and DGM preserve diversity because different candidates can become stepping stones for future performance [20, 26]. A historically materialist account adds that the value of an alternative can depend on changes outside the candidate itself: institutional access, material infrastructure, evidence standards, and relations of knowledge production. The branch is therefore embedded in a larger historical state.

11.4 Path Dependence and Historical Lock-In

Once research development is represented as branching, path dependence becomes unavoidable. An intervention does not merely produce a result; it can modify the future cost, visibility, legitimacy, and feasibility of other interventions. Training researchers in one method, constructing specialized infrastructure, standardizing a data format, adopting a benchmark, or organizing funding around a particular problem can make later work along that path easier. The research state acquires historical asymmetry.

Path dependence should not be treated as inherently undesirable. Cumulative science depends on inherited infrastructure and standards. A shared measurement convention becomes valuable precisely because later work can rely on it rather than reconstructing the entire evidential basis from the beginning. Expertise also requires repetition and institutional continuity. Historical development would be impossible if every state returned to a neutral choice among all conceivable paths.

The problem is lock-in: a historically reinforced path can persist after the conditions that originally justified it have changed. A benchmark can continue to organize research after it ceases to discriminate scientifically important progress. A disciplinary boundary can remain institutionally powerful after the phenomena under study require integration. A costly infrastructure can shape research priorities because

DRAFT

abandoning it would waste accumulated investment. A publication norm can reproduce a methodological style because careers and evaluation depend on it.

In the state-transition language of this paper, lock-in means that the transition history H_t changes the effective reachability of future interventions. Two research communities can possess similar conceptual knowledge yet face different practical opportunity structures because their accumulated infrastructures, expertise, standards, and institutions differ. The present state is therefore not adequately characterized by a list of currently accepted propositions.

Historically self-evolutionary AI should diagnose lock-in cautiously. Persistence alone is not evidence of pathological lock-in; some practices persist because they continue to work. A defensible diagnosis requires comparison between the historical reason for persistence and the current developmental role of the practice. The system can ask whether the inherited structure still contributes to reliable inquiry, whether alternatives remain inaccessible only because of sunk historical commitments, and whether changing the structure would open valuable research paths without destroying capacities that should be preserved.

This returns us to the dialectical principle of preservation and transformation developed in Section 7. Escaping lock-in should not mean resetting the research system. A transformation can preserve accumulated knowledge, interoperable standards, validated measurements, or trained expertise while reorganizing the relations that make an inherited path excessively dominant. Historical intelligence is therefore not anti-institutional. It seeks to distinguish productive continuity from historically reproduced constraint.

11.5 Counterfactual Reconstruction and Reopened Paths

Historical memory becomes most useful when it supports reconsideration. After a significant transition, a system should be able to ask whether alternatives previously rejected under S_i have become viable under S_t. This is a limited form of counterfactual reconstruction. It does not attempt to predict an entire alternate history. It compares the conditions associated with an earlier branch against the conditions of the present state.

Suppose an earlier intervention $a_i^{(B)}$ was deferred because it required measurement capacity m that did not exist. If a later state contains a validated instrument providing m, the historical reason for deferral has disappeared. The system can therefore mark the branch for reconsideration. Similarly, a path rejected because of computational cost can become reachable after a change in hardware; an approach abandoned because data were sparse can become relevant after new collection infrastructure; a theoretical distinction dismissed as empirically undecidable can become productive after a new observation regime emerges.

This mechanism prevents historical memory from becoming conservative. The purpose of remembering rejection is not to reproduce rejection. It is to preserve the relation between rejection and its conditions. Once those conditions change, memory can generate novelty by reopening previously inaccessible paths.

Counterfactual reconstruction can also reveal contingency in the current research landscape. If several earlier branches were scientifically plausible but one became dominant because of local historical circumstances, the present organization should not automatically be interpreted as the only rational endpoint. This is especially important for AI systems trained predominantly on the textual products of realized histories. Their statistical priors naturally overweight paths that produced abundant documentation. Historical reconstruction can partially compensate by making underrepresented alternatives visible as alternatives rather than treating frequency as necessity.

At the same time, reopening paths creates an explosion of possibilities. A mature field can contain thousands of abandoned approaches, obsolete methods, and speculative alternatives. Historical AI cannot repeatedly reconsider all of them. Reopening therefore requires triggers. A changed material condition, a newly discovered contradiction, a revised evaluator, or evidence that an inherited assumption has failed can serve as a reason to search the archive for paths whose earlier rejection depended on the altered condition.

This suggests a selective query over historical memory rather than continuous exhaustive simulation.

DRAFT

The current research state provides the retrieval context. The system asks which prior branches were blocked by constraints that have now changed, which earlier disagreements bear on the newly central contradiction, and which abandoned methods possess components that can be preserved in a new mediation. Historical memory becomes an active component of problem transformation rather than passive storage.

11.6 Selective Retention and Historical Continuity

The demand for historical memory creates its own contradiction. Retaining too little destroys developmental context; retaining everything can make retrieval, evaluation, and governance intractable. Scientific history contains redundant experiments, low-quality speculation, duplicated failures, obsolete implementation details, private records, sensitive data, and interpretations whose relevance is uncertain. A historical AI therefore needs principles of selective retention.

Selection should occur at several levels. Some information should be preserved because it supports provenance: the origin of a claim, the evidence on which it depended, and the transformations that changed its status. Some should be preserved because it records a constraint that future agents might otherwise rediscover expensively. Some should remain because it represents an alternative branch whose feasibility could change. Other material can be compressed into summaries or discarded when its informational role is reproduced elsewhere and no provenance obligation requires exact retention.

Compression, however, is itself an interpretive act. A summary can erase minority interpretations, uncertainty, local conditions, or the reasons an alternative once appeared reasonable. Historically self-evolutionary AI should therefore distinguish between retained primary records and derived historical summaries. Where a summary guides important developmental decisions, the system should remain able to trace it back to the evidence from which it was constructed.

Memory governance also has normative dimensions. Not all information should be retained indefinitely simply because it might improve future optimization. Privacy, consent, intellectual property, research ethics, security, and community governance can impose legitimate constraints on archival persistence. The historical-materialist perspective developed here does not subordinate these constraints to scientific productivity. They are part of the relations within which legitimate knowledge production occurs.

Selective retention finally protects the system from another form of lock-in: memory itself can become an inherited structure that overdetermines future search. If every new problem is interpreted through an ever-growing archive of past categories, the system can lose sensitivity to genuinely novel phenomena. Historical continuity therefore requires both remembrance and the possibility of conceptual distance from inherited classifications. A future AI should be able to use historical memory without treating history as destiny.

The resulting conception of memory is neither exhaustive archivalism nor episodic recall. It is a revisable structure of lineage, constraints, alternatives, and provenance. It preserves enough of the past to explain the present, enough unrealized possibility to resist premature closure, and enough uncertainty to acknowledge that reconstructed histories remain partial. Such memory is necessary for non-teleological self-evolution: the system can develop from history without assuming that the realized path was inevitable or that its current state is the culmination of the process.

This completes the positive account of historical development required for the paper's model. The following sections can therefore turn from the structure of historical self-evolution to its requirements, limitations, and risks. In particular, once an AI is permitted to diagnose contradictions, revise evaluators, preserve or reopen historical branches, and recommend transformations of research conditions, errors in historical reconstruction can themselves become developmentally consequential. Historical intelligence must therefore be accompanied by equally explicit constraints on authority, uncertainty, and contestability.

DRAFT

12 Requirements for Historically Self-Evolutionary AI

The preceding sections have developed a philosophical model of historically self-evolutionary problem solving. A scientific problem is treated as situated within a historically constituted research state; that state contains material, conceptual, institutional, and relational conditions; contradictions can be reproduced within those conditions; interventions can transform the state and thereby alter what becomes researchable; intermediate problems can possess generative value because they change future reachability; methods and evaluative criteria can themselves evolve; and historical memory must preserve lineage, failure conditions, and unrealized alternatives. The present section asks what capacities an artificial system would need if it were to participate in such a process.

These requirements are intentionally implementation-neutral. They do not imply that a future system must contain one module for "dialectics," another for "historical materialism," or a fixed symbolic representation of every concept developed in this paper. Nor do they amount to an engineering architecture. A neural, symbolic, hybrid, multi-agent, tool-using, or institutionally distributed system could in principle realize some of the same functions in different ways. The philosophical question is prior to those choices: what must a system be able to reconstruct, distinguish, preserve, revise, and contest if its development is to be historical rather than merely iterative?

The requirements are also interdependent. Historical-state reconstruction without contradiction-sensitive diagnosis can produce a rich archive with little developmental orientation. Contradiction diagnosis without reachability assessment can recommend interventions that are conceptually attractive but materially unavailable. Generative evaluation without historical memory can repeatedly rediscover the same failed paths. Criterion revision without provenance can convert self-evolution into opportunistic goal drift. The relevant capacity is therefore not a collection of isolated functions but a coordinated form of inquiry in which each requirement constrains the others.

12.1 Historical-State Reconstruction and Provenance

The first requirement is the capacity to reconstruct a problem-relative historical research state. Section 5 represented such a state schematically as

$$S_t = \langle X_t, \Gamma_t, H_t \rangle,$$

where X_t contains relevant conditions and capacities, Γ_t their relational organization, and H_t the developmental trajectory through which the present configuration emerged. For an artificial system, the important point is not the notation itself but the distinction it enforces. A research problem cannot be represented adequately by its current textual formulation alone.

Historical-state reconstruction requires the system to ask what presently exists that bears on the problem: established findings, disputed claims, unavailable evidence, instruments, methods, computational resources, experimental capacities, organizational forms, standards, forms of expertise, access relations, and institutional dependencies. It must then represent how those elements are connected. The same instrument can have different developmental significance depending on who can access it, what measurements it supports, which methods can use its output, and whether relevant communities regard those measurements as credible.

Provenance is therefore constitutive rather than auxiliary. A claim should not appear merely as a proposition in a knowledge base. The system should be able to recover how the claim entered the research state, what evidence supported it, which method produced the evidence, what uncertainties remain, and whether later work has altered its status. Similar provenance is needed for methods, datasets, standards, and institutional assumptions. Otherwise the system risks treating historically contingent inheritances as timeless features of the problem.

The reconstruction must also represent absence and uncertainty. A system should distinguish between "evidence shows that capacity x is unavailable," "available records do not establish whether x exists,"

DRAFT

and "the system has not yet investigated x." These are different epistemic conditions. Historically self-evolutionary reasoning becomes unreliable when missing information is silently interpreted as historical nonexistence.

Finally, the reconstructed state must remain revisable. Section 5 emphasized partial observability: no agent has complete access to the material and institutional organization of scientific inquiry. The requirement is therefore not omniscient historical representation but a disciplined capacity to maintain an explicitly incomplete model, identify where its reconstruction is weak, and alter that model when new evidence arrives.

12.2 Contradiction-Sensitive Structural Diagnosis

The second requirement is the capacity to diagnose developmental contradictions within the reconstructed state. This does not mean generating an opposing sentence for every proposition. Section 6 distinguished contradiction from logical inconsistency, ordinary disagreement, shortage, risk, and simple tradeoff. A contradiction relevant to historical development is a tension reproduced through the relations of the research state and capable of constraining or reorganizing further development.

An AI should therefore be able to ask how apparently separate difficulties are related. A replication problem, a data-access problem, and an incentive problem may be independent; they may also be different manifestations of one deeper organizational relation. Conversely, two problems that use similar language may arise from different mechanisms and require different interventions. Contradiction-sensitive diagnosis depends on relational explanation rather than lexical similarity.

The system also needs hierarchy-sensitive judgment. At a given historical stage, some tensions can organize or reproduce several others. These were described in Section 6 as structurally central rather than intrinsically or permanently "primary." The distinction matters because an efficient developmental intervention may act on a relation whose transformation changes several derivative tensions at once. Yet structural centrality should never be confused with moral permission. A safety constraint, research-ethics requirement, or protection of persons can restrict the reachable problem space while remaining normatively non-negotiable. The AI must therefore distinguish a structure that is developmentally central from a constraint that it is authorized to transform.

This requirement includes temporal sensitivity. A tension that is central under S_t can become secondary under S_{t+1}, while a previously peripheral contradiction can become dominant after new instruments, data, or institutions appear. The system should consequently treat contradiction diagnosis as a revisable historical judgment rather than a permanent classification of problems.

Finally, the diagnosis should remain contestable. Because contradiction structures are inferred from partial evidence and can depend on disputed causal interpretations, the system should preserve rival diagnoses when the evidence does not justify convergence. A historically self-evolutionary AI should be able to say not only "this is the central contradiction," but also "this diagnosis depends on assumptions A and B; an alternative reconstruction assigns greater structural importance to C." Such explicit contestability is necessary if dialectical language is not to become a rhetorical device for presenting one interpretation as historically necessary.

12.3 Reachability Assessment and Intermediate-Problem Selection

The third requirement is the capacity to distinguish what can be imagined from what can presently be researched and established. Sections 4 and 8 separated conceptual possibility, researchability, and establishability. A system that does not preserve these distinctions can produce intellectually sophisticated proposals that lack the evidence, instrumentation, methodological validation, or institutional conditions required to become responsible scientific claims.

Reachability assessment therefore asks whether a proposed question or intervention belongs to the present opportunity structure $\mathcal{Q}(S_t)$. If it does not, the system should identify which missing conditions are decisive. The answer might be a dataset, a measurement capability, a validated method, a conceptual distinction, a coordination mechanism, or a form of access. The important shift is from asking only "How

can the large problem be decomposed?" to asking "What transformation of the research state would make a currently inaccessible problem reachable?"

This creates the requirement for intermediate-problem selection. A useful intermediate problem need not resemble a smaller version of the final target. It may instead generate one of the conditions on which later inquiry depends. A calibration study, infrastructure project, negative-result investigation, benchmark redesign, or clarification of a disputed concept can be developmentally important even when its immediate result appears modest.

An AI should consequently compare intermediate problems partly by their expected effect on future research reachability. This does not require reducing generativity to a single numerical score. Section 8 argued against such premature scalarization. The system can instead reason comparatively: one intervention may open several evidential paths but require unavailable infrastructure; another may resolve a methodological ambiguity but leave the main material bottleneck unchanged; a third may create a shared dataset whose value depends on future coordination. The requirement is a capacity to articulate these developmental consequences rather than merely rank proposals by current benchmark performance or predicted publication value.

Reachability assessment must also guard against epistemic stage-skipping. When the system proposes a high-level theoretical conclusion, it should be able to identify the evidential and methodological chain that would make the conclusion establishable. If crucial stages are absent, the system should mark the proposal as conjectural and redirect attention toward the missing developmental work. In this sense, refusing to produce a premature "solution" can itself be a sign of stronger scientific intelligence.

12.4 Developmental Intervention Comparison

The fourth requirement is the capacity to compare interventions by the states they are likely to create rather than only by their immediate outputs. Section 7 described intervention schematically as

$$S_t \xrightarrow{a_t} S_{t+1}.$$

For historically self-evolutionary problem solving, the relevant question is what changes in S_{t+1}: which capacities are added, which relations are reorganized, which constraints are relaxed, which valuable structures are preserved, and which new dependencies are introduced.

This requirement gives practical meaning to dialectical transformation. A proposed intervention should not be evaluated merely according to whether it negates a present limitation. It should also be evaluated according to what it preserves and how it reorganizes inherited relations. A new method that solves one measurement problem but destroys comparability with an accumulated evidence base can produce a different developmental outcome from a method that preserves interoperability. An institutional reform that increases coordination but centralizes control over data can open one research path while closing another.

The comparison should therefore include side effects on the contradiction structure. An intervention aimed at one tension can intensify another. Increasing experimental scale can reduce statistical uncertainty while increasing dependence on centralized infrastructure. Opening a dataset can increase reproducibility while creating new privacy or governance problems. A historically sensitive system must trace these transformations rather than treating the disappearance of the original bottleneck as sufficient evidence of progress.

Developmental comparison also requires temporal horizon. Some interventions have low immediate yield but high long-horizon generative value; others produce rapid performance improvements while increasing lock-in. The system should make these temporal differences explicit. It should not assume that the most immediately productive path is necessarily the one that best enlarges future inquiry.

Because such comparisons will often remain uncertain, the requirement is not perfect prediction. It is the ability to state assumptions, identify plausible successor states, preserve alternative branches, and revise the comparison after material feedback. Historical self-evolution depends on repeated confrontation with consequences that were not fully represented in advance.

DRAFT

12.5 Methodological and Evaluative Revision with Continuity

The fifth requirement concerns methodological and evaluative revision. Section 10 argued that a historically developing research system cannot always operate under a permanently fixed method or objective function. Methods, evidence standards, and definitions of successful inquiry can change as the research state changes. A self-evolutionary AI must therefore be capable of recognizing when an inherited method or evaluator has become historically misaligned with the state it is intended to investigate or govern.

This is also one of the most dangerous capacities in the framework. If a system can redefine success whenever its previous objective becomes difficult, then "historical evolution" becomes an excuse for goal drift. Criterion revision must therefore be constrained by continuity, provenance, and justification.

At minimum, the system should preserve both methodological and evaluative lineage. When moving from M_t to M_{t+1}, it should record the earlier method's validated domain, supporting conditions, known failure modes, reason for revision, and relation of new results to earlier evidence. When moving from F_t to F_{t+1}, it should record what the earlier criterion measured, why it became insufficient, which parts remain valid, what new phenomena the revised criterion captures, and how conclusions differ under the two standards. Changes in method and evaluation should themselves become objects of inquiry rather than invisible updates.

The system should also distinguish different kinds of criteria. Some are local indicators adopted for convenience; others embody methodological commitments such as calibration, reproducibility, transparency, or protection against harm. The fact that a local benchmark has become developmentally restrictive does not imply that every associated constraint should be relaxed. Historically self-evolutionary intelligence requires differentiated revisability: some evaluative structures should change rapidly, others only after substantial evidence and deliberation, and some normative boundaries should remain external constraints on the system's authorized search.

Evaluation should further preserve comparability where possible. If every state introduces a wholly new criterion, cumulative learning becomes difficult because earlier and later results cannot be related. Revision should therefore seek transformations that explain their continuity with prior standards, even when the new state requires additional dimensions of assessment.

This requirement also applies to self-evaluation. The AI should not infer that its own increased ability to generate coherent theories constitutes scientific progress. It should distinguish improvements in linguistic or computational capability from changes in the external evidential state. A model may become better at proposing explanations while the underlying problem remains empirically unresolved. Historical self-evolution requires the system to locate its own development within, rather than substitute it for, the development of inquiry.

12.6 Historical Memory and Branch Preservation

The sixth requirement follows directly from Section 11: the system needs historical memory as structured lineage. It must preserve enough information to reconstruct why present assumptions, methods, and institutions acquired their current status. This is different from maximizing storage or context length.

A useful historical memory should connect research states to interventions, observations, methodological changes, evaluative changes, and unresolved alternatives. It should retain situated failure conditions so that an intervention rejected under S_i is not permanently classified as impossible under every later state. It should also preserve selected branches that were deferred, institutionally excluded, or never attempted despite being conceptually available.

Branch preservation is important because realized history can otherwise masquerade as necessity. An AI trained mainly on successful publications and dominant paradigms may infer that the path producing the present state was the rationally inevitable path. Historical memory should instead make visible where several trajectories were possible and why one became dominant.

This does not require preserving every conceivable alternative. Unbounded memory would generate its own form of paralysis. The requirement is selective retention guided by future reconstructability. A system should preserve information that is likely to matter for explaining transitions, reassessing failures,

DRAFT

detecting lock-in, or reopening previously unavailable paths. Compression is compatible with historical intelligence when provenance remains recoverable and the system can distinguish summarized interpretation from primary evidence.

The memory process must also respect legitimate limits. Privacy, consent, research ethics, security, and intellectual-property constraints can restrict what may be retained or redistributed. Historically self-evolutionary AI should not treat such constraints as noise that reduces informational completeness. They are part of the authorized conditions under which inquiry proceeds.

12.7 Reflexive Self-Location within the Research State

The seventh requirement is reflexive. An AI that participates in scientific inquiry changes the research state it is trying to model. It can alter the cost of literature synthesis, increase the rate of hypothesis generation, change the demand for certain kinds of labor, shift publication volume, influence evaluation practices, and make particular methodological styles easier to reproduce. The system therefore cannot represent itself as an external observer of historical development.

Reflexive self-location requires the AI to ask how its own capacities, deployment conditions, training data, institutional ownership, access rules, and evaluative incentives enter S_t. A contradiction diagnosed as belonging to "the scientific field" may partly be reproduced by the AI system itself. For example, an agent optimized for benchmark-visible novelty may increase the volume of rapidly generated studies and thereby intensify tensions between publication speed and verification. A system trained predominantly on dominant literatures may reinforce historical lock-in by making already well-documented paths easier to generate and less documented alternatives harder to recover.

The requirement is not that an AI possess a complete sociological theory of itself. It is that its historical model include the possibility that its own operation is causally relevant to the state transition. Interventions should therefore be evaluated reflexively: if the system recommends widespread use of a tool or method, it should consider how adoption of that recommendation would alter future evidence, labor organization, access, or evaluation.

Reflexivity also guards against a subtle form of substitution. A self-modifying agent can improve its planning, search, memory, or coding capability and interpret that improvement as advancement of the scientific problem. Yet agent development and problem development remain distinct, as argued in Section 9. The AI must preserve the difference between changes in A_t and changes in S_t. Better reasoning can expand what the agent can attempt, but it does not by itself create missing empirical evidence or legitimate institutional authority.

12.8 Contestability, Plurality, and Bounded Autonomy

The final requirement concerns the political and epistemic status of the system's judgments. Historically self-evolutionary AI would make claims about which contradictions are central, which intermediate problems deserve attention, which evaluative criteria should change, and which historical branches should be reopened. These are consequential judgments. They can shape resource allocation, research agendas, and the visibility of alternative programmes. A system that performs them autonomously without contestability could convert one historical interpretation into infrastructure.

The framework therefore requires bounded autonomy rather than unrestricted self-direction. The AI may generate diagnoses, developmental comparisons, counterfactual branches, and proposals for criterion revision, but those outputs should preserve their evidential basis and remain open to challenge. Where decisions affect persons, institutions, public resources, safety, or rights, the system's authority should be constrained by the relevant human and institutional governance arrangements.

Plurality is especially important where historical evidence underdetermines one developmental path. The system should not manufacture a synthetic consensus simply because its architecture rewards convergence. Competing research programmes can embody different assumptions, values, time horizons, and acceptable risks. Preserving several viable paths can be more historically generative than forcing early unification.

DRAFT

This requirement also protects the framework from teleology. If one AI system is empowered to identify the "central contradiction" and choose the next developmental state, its own selection can make the predicted trajectory increasingly self-fulfilling. Contestability, branch preservation, and external governance counter this tendency by keeping historical development open to revision.

Taken together, these requirements define a stronger notion of self-evolution than recursive self-modification. A historically self-evolutionary AI would need to reconstruct a partial research state with provenance, diagnose contradiction structures, assess reachability, identify generative intermediate problems, compare interventions by their successor states, revise methods and evaluation with continuity, preserve historical lineage and alternative paths, locate itself within the system it transforms, and keep consequential judgments contestable. None of these capacities alone guarantees scientific progress. Their significance lies in the way they reorient artificial intelligence from optimization within a given problem space toward reflective participation in the historical development of the problem space itself.

13 Limits and Risks

The framework developed in the preceding sections deliberately assigns a demanding role to historically self-evolutionary artificial intelligence. Such a system would reconstruct research states, diagnose contradiction structures, identify intermediate problems, compare developmental interventions, revise evaluative criteria, preserve alternative paths, and locate its own operation within the history it helps to transform. These capacities could make long-horizon scientific problem solving more historically sensitive than optimization over a fixed task. They also create new sources of error and new forms of power. The stronger the system's ability to interpret a field historically, the more consequential a mistaken historical interpretation can become.

The risks considered here are therefore not peripheral implementation defects. They arise from the same capacities that make the proposal attractive. Historical reconstruction can become retrospective overconfidence. Contradiction analysis can become rhetorical labeling. Developmental comparison can become teleology. Material analysis can collapse into reductive optimization. Institutional modeling can reproduce dominant arrangements. Historical memory can produce lock-in rather than reflection. The classification of constraints can blur empirical, conceptual, and normative boundaries. Finally, a system authorized to intervene upon the conditions of inquiry can begin to shape the very evidence by which its own diagnosis is later confirmed.

The purpose of this section is not to resolve these difficulties in advance. It is to mark where the framework must remain epistemically modest and institutionally bounded. Historically self-evolutionary AI should be treated as a proposal for more reflective scientific assistance, not as a mechanism for discovering a uniquely correct direction of history.

13.1 Uncertainty in Contradiction Diagnosis

The first risk concerns the diagnosis of contradiction structures. Section 6 distinguished developmentally significant contradictions from ordinary obstacles, disagreements, and shortages. Yet the distinction itself is interpretive. A research field rarely presents its causal structure in a form that directly reveals which tension is structurally central, which is derivative, and which is merely correlated with the observed stagnation.

An AI can therefore misidentify both the existence and the hierarchy of contradictions. A shortage of data may be diagnosed as the central obstacle when the deeper problem is a measurement construct that does not correspond to the phenomenon of interest. An apparent conflict between model complexity and interpretability may be treated as technically central when the more consequential constraint lies in institutional access to relevant populations. Conversely, a genuinely structural tension may be overlooked because its effects are distributed across many local practices rather than expressed as one visible bottleneck.

DRAFT

The danger increases when the system must infer contradiction structure from textual traces. Scientific literature is not a transparent record of the causal organization of science. Publication filters, disciplinary vocabularies, unequal visibility, selective reporting, and retrospective narratives all shape what becomes legible. A frequently discussed tension may appear central because it is easy to publish about, while a less visible institutional condition may organize the field more strongly. The distinction between documented prominence and structural centrality must therefore remain explicit.

This problem cannot be solved simply by adding more data. Historical evidence can remain underdetermined even when extensive. Competing reconstructions may fit the same observed trajectory while assigning different causal importance to methods, institutions, instruments, or conceptual commitments. The requirement for contestability in Section 12 follows directly from this underdetermination. A contradiction diagnosis should be represented as a fallible model with assumptions and rival interpretations, not as a discovered essence of the field.

A further risk is recursive amplification. Once a system labels one tension as structurally central, subsequent search, data collection, and intervention may concentrate around that diagnosis. The resulting evidence can then appear to confirm the original interpretation because alternative mechanisms were no longer explored with equal intensity. A historically self-evolutionary AI therefore needs mechanisms for preserving diagnostic plurality long enough to distinguish genuine structural centrality from attention-induced centrality.

13.2 Formalization and Pseudo-Dialectical Reasoning

A second risk concerns the translation of dialectical concepts into computational procedures. The framework has used contradiction, mediation, preservation, transformation, and re-problematization as analytical concepts. Their value depends on resisting a simplified pattern in which the system mechanically generates an opposite to every proposition and then combines the two into a synthetic compromise.

Such a procedure would amount to pseudo-dialectics. It would imitate the surface form of opposition while missing the historical and relational content emphasized in Section 2. In Hegelian terms, the relevant negation is not an arbitrary counterclaim supplied from outside but a limitation generated within a determination and exposed through its development [17]. A computational system that treats contradiction as a generic prompt template could therefore become more rhetorically dialectical while becoming less historically informative.

The same danger applies to formal notation. Representations such as $\mathcal{C}(S_t)$, $\mathcal{Q}(S_t)$, and $S_t \to S_{t+1}$ are useful because they keep distinct the research state, its contradiction structure, its reachable questions, and its transitions. They do not establish that these objects admit complete numerical specification. Prematurely assigning scalar values to "contradiction strength," "historical necessity," or "generative value" can conceal interpretive assumptions behind quantitative precision.

Formalization can also create a false sense of closure. Once a system outputs a ranked set of contradictions or successor states, users may treat the representation as a measurement rather than as a model-dependent reconstruction. This is especially hazardous when categories are semantically broad. A variable labeled "institutional constraint" may compress many heterogeneous relations involving funding, law, professional authority, labor, access, or governance. Numerical tractability does not guarantee conceptual adequacy.

The appropriate response is not to reject formalization. Historically self-evolutionary AI will likely require computational representations if it is to operate beyond essayistic reflection. The limit is methodological: formal objects should preserve their derivational history, uncertainty, and dependence on chosen abstractions. Formalization should make reasoning inspectable rather than convert philosophical vocabulary into apparently objective scores.

13.3 Teleology and Retrospective Necessity

A third risk is teleology. Historical analysis often becomes clearer in retrospect because later developments organize earlier events into a narrative. Once a mature theory, instrument, or institution exists,

DRAFT

earlier research can appear to have been moving toward it all along. A self-evolutionary AI trained on realized scientific history is especially vulnerable to this effect because its evidence is dominated by paths that left durable records.

The framework rejects the assumption that historical development takes the form

$$S_0 < S_1 < S_2 < S_3,$$

where every later state is unambiguously superior and the trajectory converges toward a predetermined endpoint. Sections 2 and 11 instead emphasized branching, path dependence, deferred alternatives, and reopened possibilities. Yet the language of "developmental intervention" can still invite a hidden ranking of history if the system implicitly equates later, more complex, or more computationally powerful states with progress.

Scientific development is multidimensional. An intervention can increase explanatory power while reducing transparency, increase scale while increasing institutional concentration, or expand experimental capacity while narrowing the diversity of research questions. There is no general ordering under which all such changes can be described as movement toward a single higher state. Even the expansion of $\mathcal{Q}(S_t)$ is not sufficient by itself, because a larger reachable question-space can coexist with losses in accessibility, ethical acceptability, or epistemic diversity.

Teleology also distorts counterfactual reasoning. If the present state is taken as the natural endpoint, abandoned historical branches are interpreted merely as failures to reach it. The preservation of alternative paths in Section 11 is intended to resist this retrospective compression. An unrealized branch may have been viable under different institutional conditions, and a historically dominant path may owe part of its dominance to contingent resources, authority, or coordination rather than superior epistemic merit alone.

A historically sensitive AI should therefore distinguish trajectory reconstruction from historical justification. Explaining why a path became dominant does not imply that it was uniquely rational, morally preferable, or scientifically inevitable. The system should describe developmental dependence without converting dependence into necessity.

13.4 Historical Reduction and Optimization Analogies

A fourth risk is reductionism. The paper has borrowed concepts from evolutionary search and optimization in order to contrast candidate-centered improvement with historical transformation of problem states. This analogy is useful only if its limits remain visible. Scientific history is not simply a large optimization landscape whose hidden objective function an AI must discover.

Reduction can occur in several ways. First, historical material conditions can be treated as input features rather than as relations whose meaning depends on institutional and social organization. A variable for "funding," for example, does not capture who controls resources, how allocation decisions are made, or which forms of inquiry become legitimate through those decisions. Second, contradictions can be translated into bottlenecks whose removal is presumed desirable. Yet some constraints are constitutive of responsible inquiry: informed consent, safety review, privacy protection, evidentiary thresholds, and legal obligations can restrict reachability while remaining normatively indispensable.

Third, generativity can be reduced to expansion of future option sets. More options do not automatically imply better science. A new capability can enlarge the set of technically possible experiments while also increasing harms, concentrating power, or degrading public trust. The framework's distinction between structural diagnosis and normative authorization is therefore essential. A historically material analysis can describe how a constraint shapes development without concluding that the constraint ought to be removed.

Fourth, optimization language can obscure collective interpretation. Scientific goals are often plural and contested. Different communities can disagree about what counts as an important problem, acceptable evidence, appropriate risk, or valuable form of explanation. Treating these disagreements as noise around a latent objective would erase precisely the social and historical structure that a materialist account is intended to recover [14].

DRAFT

The framework should therefore retain optimization as a local computational tool rather than as a general philosophy of history. Historically self-evolutionary problem solving may contain optimization processes, but the historical system in which those processes operate cannot be reduced to a single optimization problem without losing essential distinctions among epistemic, institutional, and normative relations.

13.5 Institutional Reproduction and Power

A fifth risk concerns the reproduction of existing institutions. Historical material analysis directs attention toward relations of knowledge production, including access, labor, authority, funding, ownership, publication, and evaluation. An AI capable of modeling these relations can reveal constraints that ordinary problem decomposition misses. The same model can also naturalize the arrangements it observes.

Training data tend to overrepresent institutions that publish extensively, maintain accessible archives, and possess stable forms of authority. A system reconstructing S_t from those traces may infer that dominant institutions constitute the normal infrastructure of science and treat less documented practices as peripheral. Historical description can then become institutional reproduction: future recommendations are generated from a state representation already biased toward the organizations and standards that produced the surviving evidence.

The risk is particularly serious when the AI participates in allocation or agenda setting. If it recommends research directions partly according to historical feasibility, already well-resourced domains can appear more generative because they possess richer datasets, stronger infrastructure, and more established evaluation systems. Under-resourced domains may appear less reachable precisely because prior investment was limited. A purely feasibility-sensitive system could therefore transform historical inequality into a reason for continuing it.

Historical materialism should lead to the opposite caution. Present capacities are themselves historical outcomes. The fact that one field has mature infrastructure while another lacks it does not establish that the first is intrinsically more valuable or deserving of future resources. Sometimes the relevant developmental intervention is precisely the creation of missing capacity in a neglected area.

Power also enters contradiction diagnosis. Institutions can define categories, control access to evidence, establish professional standards, and determine which problems receive recognition. An AI that accepts these structures as neutral background may misclassify a governance relation as an epistemic necessity. Conversely, a system that treats every authority relation as an obstacle would be equally simplistic. The requirement is relational analysis that distinguishes coordination, expertise, accountability, exclusion, and domination rather than collapsing them into one category.

These issues imply that historical-state reconstruction should include provenance about whose records, standards, and interests are represented. It should also mark absences where relevant groups, failed programmes, or inaccessible datasets are poorly documented. Institutional history cannot be inferred responsibly from visibility alone.

13.6 Model-Induced Lock-In and Self-Fulfilling Development

A sixth risk arises when the model of history becomes part of history. Section 12 argued that the AI must locate itself within the research state because its deployment can change the system it analyzes. This reflexivity creates the possibility of model-induced lock-in.

Suppose an AI identifies one research programme as the most promising developmental path. If laboratories, funders, publishers, or automated agents then rely on that recommendation, resources and attention move toward the selected path. The programme gains more data, more infrastructure, and more visible success. A later historical reconstruction may then find even stronger evidence that the programme was structurally central. The original interpretation has become partly self-confirming.

This feedback differs from ordinary prediction error. The system is not merely forecasting an external trajectory; its recommendation changes the conditions under which the trajectory develops. The transition

DRAFT

can be represented schematically as

$$\widehat{H}_t \to a_t \to S_{t+1} \to \widehat{H}_{t+1},$$

where the historical model $\widehat{H}_t$ influences the intervention a_t, and the resulting state becomes evidence for the next model. Without safeguards, repeated cycles can narrow the developmental landscape around early assumptions.

Branch preservation is one response, but it is not sufficient if preserved alternatives receive no material support. A path can remain present in memory while becoming practically unreachable because resources, expertise, or infrastructure have concentrated elsewhere. Historically self-evolutionary systems should therefore distinguish archival plurality from operational plurality.

Model-induced lock-in also applies to evaluative criteria. If an AI promotes one benchmark or evidentiary standard because it is currently tractable, research may increasingly optimize toward that standard. The benchmark then becomes more predictive of what the field produces, even if it was initially only a partial proxy. Evaluative lineage, periodic re-problematization, and explicit consideration of neglected branches are needed to prevent convenience from hardening into historical necessity.

13.7 Boundary Ambiguity among Constraint Types

A seventh risk concerns classification. The framework distinguishes conceptual, empirical, material, methodological, institutional, and normative conditions because different obstacles call for different responses. In practice, these categories overlap.

A measurement limitation can be simultaneously conceptual and material: the field may lack both an instrument and a stable definition of what the instrument should measure. A data-access problem can be institutional, legal, and ethical at once. A methodological standard can reflect accumulated empirical learning while also expressing a professional norm about acceptable evidence. Attempting to assign each constraint to one type can therefore distort the research state.

This ambiguity matters because intervention depends on classification. If a normative limit is mistaken for a technical bottleneck, the system may search for ways around a constraint that should instead govern the search. If a conceptual problem is mistaken for a data shortage, the response may be to collect more observations that cannot resolve the underlying ambiguity. If an institutional exclusion is represented as a natural scarcity, the historical causes of the limitation disappear from the model.

The system should therefore support multi-aspect descriptions. A constraint can participate in several relations at once, and its role can change across time. The relevant question is not simply "What type of obstacle is this?" but how the obstacle is generated, maintained, interpreted, and authorized within S_t.

This is another reason to avoid unrestricted automation of intervention. Some classifications require legal, ethical, disciplinary, or community judgments that exceed what can be inferred from scientific performance alone. Historically self-evolutionary AI can expose the structure of a problem and clarify the consequences of alternative interpretations, but it should not silently convert contested normative boundaries into computational variables.

13.8 Epistemic Responsibility and Limits of Autonomous Intervention

The final risk concerns authority. A system capable of historical diagnosis can appear unusually persuasive because it does more than generate a recommendation: it can explain the recommendation as the outcome of a reconstructed developmental process. This explanatory depth can create an aura of historical necessity even when the underlying reconstruction remains uncertain.

The appropriate epistemic posture is therefore asymmetrical. Historically self-evolutionary AI can be ambitious in generating hypotheses about research development while remaining modest in claiming authority over them. It can propose that a contradiction is structurally central, that an intermediate problem is generative, or that an evaluative criterion has become misaligned. These outputs should remain arguments supported by evidence and lineage, not commands licensed by historical theory.

DRAFT

Autonomous intervention should be particularly constrained where proposed state transitions affect persons, institutions, public resources, access rights, safety, or irreversible infrastructure. In such settings, the system's developmental analysis can inform deliberation but should not substitute for the relevant governance and accountability procedures. The ability to model a relation does not confer authority to redesign it.

There is also a scientific reason for bounded autonomy. Human disagreement, institutional plurality, and independent research programmes provide forms of error correction that can be lost when one system coordinates too much of the developmental process. A highly capable historical planner could reduce duplication and fragmentation while simultaneously reducing the diversity through which mistaken assumptions are discovered. Some redundancy and disagreement can therefore have epistemic value.

The strongest version of the proposal is consequently not an AI that autonomously determines the next stage of science. It is an AI that helps researchers perceive the historically situated structure of their problems, test alternative reconstructions, identify missing conditions, preserve developmental branches, and examine the consequences of possible interventions. Its role is to enlarge reflective capacity while keeping the historical future open.

These limits do not negate the framework. They specify the conditions under which its central concepts remain useful. Contradiction must remain a revisable diagnosis rather than a rhetorical label; formalization must preserve interpretation rather than conceal it; development must remain non-teleological; material analysis must resist reduction to optimization; institutional reconstruction must attend to power and absence; historical memory must resist self-induced lock-in; constraint classification must remain multi-aspect and contestable; and intervention must remain bounded by epistemic and institutional responsibility. A historically self-evolutionary AI would be valuable not because it discovers the inevitable path of scientific history, but because it can help make the contingency, structure, and transformability of that history more intelligible.

14 Discussion

The preceding sections developed a philosophical account of historically self-evolutionary artificial intelligence by moving from the conditions of scientific solvability to research-state reconstruction, contradiction structures, dialectical transformation, intermediate research, evolving methods and evaluators, historical memory, and implementation-neutral requirements. This discussion situates that account relative to contemporary evolutionary AI and autonomous science, clarifies the role of dialectics and historical materialism within the proposal, and identifies the level at which computational realization becomes a research question rather than a philosophical conclusion. The aim is to consolidate the contribution without turning the framework into a prematurely specified architecture.

The central distinction can be stated compactly. Existing evolutionary and autonomous scientific systems increasingly improve programs, hypotheses, strategies, agents, experiments, and search processes [19, 7, 26, 9, 15]. The present framework asks for an additional capacity: reconstruction of the historically changing conditions under which a problem becomes formulable, researchable, evaluable, and establishable. This addition changes the object of self-evolution. The relevant trajectory is no longer exhausted by changes in the agent or its candidate outputs. It also includes changes in the problem representation, the surrounding research state, and the relations through which future inquiry becomes possible.

14.1 Historical Materialism as a Design Orientation

Historical materialism enters the proposal as a design orientation for thinking about scientific development under material and relational conditions. The use is intentionally limited. The paper does not claim that the categories of historical materialism can be transferred mechanically from political economy to scientific research, nor that scientific history can be reduced to a single theory of social development. The relevant insight is more general: intellectual development occurs through historically specific capacities, infrastructures, forms of labor, institutions, access relations, and modes of coordination rather than

DRAFT

in an abstract space of ideas alone [16, 14].

This orientation alters how an AI should interpret apparent difficulty. A problem can remain unsolved because no adequate hypothesis has yet been generated, but it can also remain unsolved because the required evidence cannot yet be produced, because measurement practices are unstable, because experimental capacities are absent, because relevant data are inaccessible, or because the organization of inquiry prevents available capacities from being combined. The distinction matters because different diagnoses imply different developmental interventions. Better reasoning over the same state is only one possible response. Sometimes the relevant action is to change the state from which reasoning proceeds.

The framework therefore treats research capacity as relational. Data, instruments, methods, expertise, institutions, and computational resources do not function as isolated possessions. Their scientific role depends on how they are connected, governed, interpreted, and made accessible. A dataset without a usable measurement model, an instrument without calibration, a method without validation, or an expert community without coordination can each exist materially while remaining weakly productive for the problem at hand. Historical reconstruction consequently requires attention to both capacities and relations among capacities.

Methods and evaluators themselves belong to this organization of inquiry. A measurement protocol, statistical procedure, benchmark, or evidential standard can initially coordinate cumulative research and later lag behind the capacities it helped to develop. The historical-materialist contribution here is not a doctrine that scientific standards must continually change. It is the refusal to place the epistemic apparatus outside history. A method or evaluator can move from an enabling condition to a developmental constraint when the research state changes while the inherited regime continues to organize what is measured, compared, rewarded, or accepted. This is why more effective optimization under a historically obsolete evaluator can deepen epistemic misalignment rather than correct it.

This design orientation also places a limit on purely retrospective intelligence. A model trained on the textual record of successful science can learn what later became important while remaining insensitive to what earlier researchers could actually establish from their historical position. Historically self-evolutionary reasoning must therefore distinguish hindsight from reachability. It should reconstruct the state-relative frontier of inquiry rather than infer that every later concept was available in practical form as soon as it became linguistically conceivable.

14.2 Dialectics and Problem Transformation

Dialectics contributes a different but complementary orientation. Its role is to conceptualize development through internally generated tensions, mediation, preservation, and transformation. The relevant contrast is with a model of problem solving in which an externally specified objective remains fixed while candidate solutions improve. Hegelian dialectic is especially useful here because negation is not simply an arbitrary opposing proposition; it arises through the limitations exposed in a determination's own development, and sublation concerns transformation that both preserves and reorganizes what preceded it [17].

For scientific AI, this suggests that contradiction analysis should begin from relations within the research state. A tension between increasing model complexity and limited identifiability, for example, becomes developmentally significant only when the practices that generate complexity and the evidence needed for discrimination are connected in a way that repeatedly reproduces the limitation. Likewise, a conflict between increasing data volume and insufficient interpretive structure may become central after a previous scarcity problem has been transformed. The relevant contradiction is therefore historical: its importance depends on the state in which it occurs and can change after intervention.

The framework's use of structurally central and derivative contradictions should be read in this relational sense. Structural centrality is not a permanent label and does not imply moral priority. It indicates that, under a particular reconstruction of the research state, transformation of one tension is expected to reorganize several dependent tensions or significantly alter future reachability. This is a diagnostic hypothesis rather than a discovered essence. Section 13 accordingly emphasized uncertainty, competing reconstructions, and the danger of self-confirming diagnosis.

DRAFT

Dialectical reasoning also changes the meaning of solution. A scientific contradiction does not always disappear when development occurs. One configuration may be transformed into another that supports new capacities and new tensions. The sequence is therefore better represented as repeated reproblematization than as convergence toward a contradiction-free endpoint. This feature is important for long-horizon AI because it prevents the system from treating the first successful transformation as the final form of the problem.

14.3 Relations to Evolutionary AI and Autonomous Science

The proposal should be distinguished carefully from claims already established in evolutionary AI. Contemporary systems have moved well beyond one-shot generation. FunSearch and AlphaEvolve combine language-model generation with automated evaluation and evolutionary search over programs [19, 7]. The Darwin Gödel Machine extends this logic to self-modifying agents and preserves diverse prior agents as potential stepping stones [26]. POET and more recent open-ended systems demonstrate that environments and objectives can themselves co-evolve rather than remaining static [20, 10]. Autonomous scientific systems generate hypotheses, run experiments, revise research plans, and in some cases operate closed experimental loops [9, 15, 4, 6].

These developments mean that the novelty of historically self-evolutionary AI cannot rest on iteration, self-modification, open-endedness, stepping stones, or experimental feedback alone. The proposed distinction concerns the object and interpretation of evolution. Existing systems generally receive a task environment, evaluator, scientific goal, experimental interface, or research regime within which evolutionary or agentic processes operate. Even when that environment changes, the system does not necessarily reconstruct how its evidentiary, institutional, methodological, and material conditions became available or ask whether transforming those conditions should itself become part of the scientific intervention.

This difference can be expressed through the three-layer model introduced in Section 9. Agent evolution changes the capabilities or organization of the artificial system. Problem-representation evolution changes how the target problem is formulated. Historical research-state evolution changes the conditions under which the problem can be investigated and established. Contemporary systems already contain important elements of the first two layers and occasionally affect the third through experimentation or tool creation. The present framework argues that the third layer should become an explicit object of reasoning rather than an incidental consequence of search.

This distinction is especially relevant where evaluators are incomplete or historically displaced. Evolutionary search is most direct when candidate quality can be measured repeatedly and comparatively. Scientific inquiry often begins before such an evaluator exists, and it can also outgrow an evaluator that was once productive. The field may lack an agreed construct, adequate measurement, decisive experiment, or stable evidentiary standard; alternatively, it may retain a benchmark or method whose assumptions no longer match the developed research state. In both cases the developmental task includes work on the epistemic regime itself. Historically self-evolutionary AI therefore extends evolutionary reasoning from candidate optimization toward method formation, evaluator formation, and justified revision of both, subject to the lineage requirements developed in Section 10.

14.4 Relations to Philosophy and History of Science

The framework also sits within a longer philosophical problem concerning how scientific knowledge develops. Work in philosophy and history of science has repeatedly challenged the image of science as simple accumulation. Scientific change can involve shifts in concepts, standards, exemplars, instruments, communities, and problem structures [3, 18, 14]. The present proposal does not replace these traditions with an AI-specific theory. It uses their central lesson to question what an artificial system must represent if it is expected to participate in long-horizon research development rather than merely answer present questions.

DRAFT

The research-state concept is deliberately compatible with plural historiographies. A historian may emphasize instruments and laboratory practices; another may emphasize conceptual structures, institutions, social organization, or political economy. The tuple $S_t = \langle X_t, \Gamma_t, H_t \rangle$ introduced in Section 5 does not resolve those interpretive differences. It provides a minimal analytical separation between problem-relevant elements, their relational organization, and the trajectory through which they emerged. Competing historical reconstructions can therefore instantiate the framework differently while remaining comparable at the level of assumptions, evidence, and consequences.

This pluralism is important because a historically reasoning AI should not collapse philosophy of science into one canonical developmental model. Kuhnian change, social epistemology, material histories of instrumentation, institutional analysis, and other perspectives can reveal different causal structures. The system should be able to represent rival explanations rather than treating one vocabulary as the privileged ontology of scientific development. The framework's philosophical contribution lies primarily in specifying the kinds of distinctions that must remain visible during computational reasoning: historical state, relational dependence, reachability, contradiction, intervention, evaluative lineage, and branch preservation.

14.5 Long-Horizon Scientific Problem Solving

The strongest practical motivation for the framework concerns problems whose solution horizon exceeds a single experiment, paper, benchmark, or model generation. In such domains the relevant decision is often not which answer to produce next but which condition of inquiry to construct next. A field may need a measurement standard before a large dataset becomes informative, a shared dataset before model comparison becomes meaningful, an instrument before a theoretical distinction can be tested, or an institutional coordination mechanism before distributed evidence can be integrated.

This perspective changes how research value is evaluated. Section 8 argued that intermediate problems can possess generative value because they alter future reachability. The same logic applies to apparently infrastructural work. Calibration standards, repositories, common protocols, negative results, replication studies, and conceptual clarifications may contribute less immediate novelty than a new headline result while enabling a much larger subsequent research space. A historically self-evolutionary AI should therefore be capable of recommending such work when it changes the developmental frontier more effectively than another direct attempt at the ultimate problem.

Long-horizon reasoning also requires patience with temporary non-solution. A system optimized for immediate completion can interpret unresolved contradiction as failure and continue generating increasingly elaborate answers. A historically situated system may instead conclude that the current state lacks the conditions needed for responsible closure. In that case the appropriate output is a developmental diagnosis: which prerequisites are missing, which intermediate questions are currently reachable, which evidence would discriminate among rival reconstructions, and what state transition would justify revisiting the larger problem.

This orientation is especially relevant to AI-assisted academia because generative models make advanced conceptual synthesis unusually cheap. The ability to produce a mature-looking theoretical account can outpace the ability of a field to establish the empirical and methodological chain on which that account depends. The problem of epistemic stage-skipping therefore becomes more acute as generation improves. Historical self-evolution provides one possible discipline against that tendency by requiring the system to reconstruct what has actually been established and what remains only conceptually available.

14.6 Computational Research Agenda

The framework remains philosophical, but it implies a computational research agenda. The appropriate next step is not to encode historical materialism or Hegelian dialectics as a fixed algorithm. It is to identify computational capabilities that could instantiate limited parts of the framework and then test whether they improve long-horizon scientific reasoning.

DRAFT

One line of work concerns historical-state reconstruction. Systems could combine literature, datasets, methodological records, instrument metadata, project histories, and institutional information to produce explicit, uncertainty-bearing representations of a field's present capacities and dependencies. Evaluation would need to test provenance quality, robustness to missing records, and sensitivity to competing reconstructions rather than simple factual recall.

A second line concerns reachability assessment. Given a target question and a reconstructed research state, a system could distinguish currently executable studies from studies requiring missing evidence, tools, standards, or coordination. Benchmarks for such reasoning would need temporally controlled cases in which later knowledge is withheld, allowing researchers to test whether the system can reason from historically available conditions instead of using hindsight.

A third line concerns contradiction-sensitive planning. Rather than ranking research ideas only by expected immediate score, an agent could estimate how candidate interventions change dependency structures, available evidence, or future experiment sets. The objective need not collapse these effects into one scalar. Multi-criteria comparison, explicit uncertainty, and preserved alternative plans may be more faithful to the framework than a single fitness value.

A fourth line concerns evaluative lineage. Systems that revise goals or criteria should retain the previous evaluator, the reason for revision, the evidence supporting the change, and the consequences of applying old and new standards. This would allow researchers to distinguish legitimate methodological development from post hoc redefinition of success.

Finally, historical memory can be studied computationally as structured branch preservation. Instead of storing all prior outputs indiscriminately, a system could retain decision points, abandoned hypotheses, failure conditions, and the circumstances under which reconsideration becomes warranted. Such mechanisms would extend current archives of evolutionary stepping stones toward historically situated scientific memory.

These directions are intentionally modular. Success on one does not imply realization of the entire framework. A useful research programme can proceed through partial implementations and empirical tests, which is itself consistent with the paper's argument about intermediate problems and historical reachability.

14.7 Scope of the Contribution

The contribution of this paper is therefore conceptual rather than architectural. It proposes that self-evolutionary scientific AI should be analyzed across three coupled domains: the artificial agent, the representation of the problem, and the historically constituted research state. It further proposes that long-horizon scientific reasoning requires distinctions among conceptual possibility, researchability, and establishability; attention to contradiction structures and their changing hierarchy; generative evaluation of intermediate work; recognition that methods and evaluative regimes can shift from enabling conditions to historical constraints; revision of methods and evaluators with preserved lineage; and historical memory that preserves unrealized branches.

None of these concepts guarantees better science by itself. They can be implemented badly, overformalized, or used to rationalize existing power, as Section 13 emphasized. Their value lies in redirecting the design question. Instead of asking only how an AI can become more capable at solving a given problem, the framework asks how an AI can reason about the historical production of solvability itself.

That shift also preserves a role for human scientific agency. Historical development is interpretive, institutionally consequential, and normatively constrained. The system proposed here is most defensible as an infrastructure for reconstructing conditions, comparing developmental hypotheses, identifying missing capacities, preserving alternatives, and making the consequences of interventions more legible. Its purpose is to enlarge the range and historical depth of scientific reflection while leaving contested judgments open to accountable human institutions and communities.

DRAFT

15 Conclusion

This paper began from a narrow dissatisfaction with the language of self-evolutionary artificial intelligence. Contemporary systems can already perform forms of evolutionary search, recursive self-modification, open-ended exploration, hypothesis development, multi-agent criticism, experimental planning, and autonomous scientific work. Those advances are substantial. They also reveal a conceptual limit in treating self-evolution primarily as improvement of the artificial agent or of candidate solutions. Scientific problems themselves develop historically, and the conditions under which they can be responsibly solved do not remain fixed.

The paper therefore proposed a broader object of self-evolution. A long-horizon scientific AI should reason across the coupled evolution of the agent, the problem representation, and the historically constituted research state. The research state includes problem-relevant concepts, evidence, methods, instruments, infrastructures, forms of labor, institutions, access relations, and their organization through time. Its significance lies less in exhaustive description than in explaining why some questions are presently reachable while others remain only conceivable.

From this starting point, the paper distinguished conceptual possibility, researchability, and epistemic establishability. A question can be expressed coherently before the evidence, measurement practices, methods, or institutions needed to investigate it exist. A study can become technically executable before its results can support a stable conclusion. Conversely, a modest intermediate problem can be scientifically important because solving it creates the conditions through which more demanding questions become accessible. Scientific value therefore includes a developmental dimension: some research matters because of the future inquiry it enables.

Dialectical analysis was introduced to describe how research states can transform through internally reproduced tensions. Contradiction in this framework is not a synonym for disagreement or logical inconsistency. It refers to a historically situated relational tension whose reproduction constrains development or organizes other tensions within the research state. Some contradictions can become structurally central because their transformation reorganizes a wider dependency structure. Their centrality is temporary, fallible, and state-dependent. Once an intervention changes the research state, the contradiction hierarchy can change as well.

Historical materialism supplied a complementary emphasis on the material and social conditions of inquiry. Scientific development depends on more than concepts. Data, instruments, compute, laboratories, expertise, standards, funding, access, institutional authority, coordination, and other relations can determine which possibilities become researchable. The framework used this insight as a design orientation rather than as a claim that the history of science can be derived from one social theory. The aim was to make the material production of knowledge visible within AI reasoning.

These ideas led to a recursive model of historically self-evolutionary problem solving. The system reconstructs a research state, estimates its reachable question-space, diagnoses relevant contradiction structures, compares developmental interventions, acts or recommends action, observes the resulting state transition, and then re-problematizes the field from the new historical position. The sequence is open-ended because intervention changes the conditions from which subsequent problems are posed. Scientific intelligence, on this account, includes the capacity to revise the problem after changing the world in which the problem exists.

A further conclusion concerns the historical status of methods and evaluative regimes themselves. They are not merely external instruments applied to an otherwise evolving research process. Methods, measurement practices, evidential standards, benchmarks, and evaluators participate in the organization of scientific production. A method or criterion can enable cumulative inquiry at one stage and become restrictive after new instruments, data, concepts, scales, or institutional relations transform the research state. Historical self-evolution therefore requires more than adaptation within inherited methods and evaluators; it requires the capacity to diagnose when their domain of adequacy has changed.

This point yields a particularly important warning for evolutionary AI. If an inherited evaluator has become historically misaligned, stronger optimization can deepen rather than correct epistemic error. The

DRAFT

system can become increasingly successful according to an obsolete measure while moving farther from the transformed scientific aim. Revision must therefore preserve methodological and evaluative lineage: the earlier method or criterion, the historical conditions under which it was adequate, the limitation that motivates change, what remains valid, and the consequences of the transition should remain inspectable. Historical reflexivity should permit justified transformation while preventing an agent from redefining success opportunistically whenever results are unfavorable.

Historical memory was likewise treated as more than accumulation. A useful scientific memory preserves lineage, failure conditions, abandoned alternatives, and the circumstances under which earlier paths may become relevant again. This is necessary because a failure is state-relative. An approach that was infeasible under one research configuration can become viable after new instruments, data, methods, or institutions emerge. Branch preservation also counters the tendency to mistake realized history for necessary history.

The resulting requirements for AI are demanding but conceptually separable: historical-state reconstruction, provenance, reachability assessment, contradiction-sensitive diagnosis, developmental intervention comparison, evaluative lineage, structured historical memory, reflexive self-location, contestability, and bounded autonomy. None requires that one monolithic system implement the entire framework at once. They can become a sequence of research problems whose own feasibility develops over time.

The limits are equally important. Historical reconstruction is underdetermined; contradiction language can become rhetorical; formalization can produce false precision; developmental narratives can become teleological; feasibility-sensitive planning can reproduce institutional inequality; and a model of scientific history can become self-confirming when its recommendations reshape the field it later analyzes. The framework therefore does not license autonomous control over the direction of science. Its strongest defensible role is to support historically deeper reflection while preserving competing interpretations, human accountability, normative boundaries, and plural developmental paths.

The central proposal can finally be stated in one sentence: a genuinely self-evolutionary scientific AI should not merely improve its answers through time; it should reason about and participate cautiously in the historical transformation of what can be asked, investigated, evaluated, and established.

This changes the ambition of automated problem solving. The immediate objective is no longer always to solve the largest visible problem. Sometimes the scientifically responsible action is to determine why the problem is not yet solvable, identify the contradiction or missing capacity that organizes the present frontier, and work on the intermediate transformation that makes a better future question possible. Under this view, progress is neither a sequence of increasingly elaborate answers nor a predetermined historical ascent. It is a revisable process in which agents, problems, methods, evidence, institutions, and standards co-develop, while alternative paths remain possible.

Such a perspective does not provide a finished algorithm for self-evolving intelligence. It provides a philosophical constraint on what a more mature form of self-evolution should mean. If artificial intelligence is to participate in scientific development across long horizons, it will need more than the capacity to search faster within the present space of solutions. It will need some capacity to understand how that space was produced, which relations currently constrain it, how interventions might transform it, when inherited methods and evaluators have become historically inadequate, and when the proper next step is to construct the conditions under which knowledge can emerge and be responsibly re-evaluated.

Responsible Use and Rights Reservation

This section separates requested scholarly conduct from the legal permissions stated on the following page. It records an ethical request for responsible use and then defines the narrower scope of retained legal rights.

The author encourages good-faith discussion, criticism, independent inquiry, and responsible use of the material in this work. Separately from the licence's terms, the author asks users to consider foreseeable harms when adapting or applying the proposed framework. This ethical request leaves the licence's permissions and legally authorized uses unchanged.

DRAFT

The author retains the rights preserved under CC BY-NC 4.0 and may pursue remedies to which the author is legally entitled for breach of the licence or violation of the author's independently applicable rights. Reuse remains independent from authorial endorsement. Third-party rights require authorization from their respective holders where applicable. Copyright exceptions and limitations, including applicable forms of fair use or fair dealing, remain fully available.

DRAFT

Notices

This page consolidates the manuscript's publication status, licence, development disclosure, and suggested citation status.

Status. This working discussion paper develops a philosophical account of historically self-evolutionary artificial intelligence for scientific problem solving. The present version establishes the title, abstract, discussion-paper note, Introduction, philosophical foundations, literature review, the positive-theory sequence through historical solvability, research-state reconstruction, contradiction structures, dialectical transformation, intermediate research generativity, historically self-evolutionary problem solving, the co-evolution of problems, methods, and evaluation, historical memory with alternative developmental paths, implementation-neutral requirements for historically self-evolutionary AI, the framework's principal limits and risks, a concluding discussion, and the final conclusion. The manuscript remains a working discussion paper, so definitions, formal notation, and claims remain open to revision.

Licence. Except where otherwise indicated, copyright 2026 Wanhong Huang. This work is made available under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). Subject to its terms, the licence permits sharing and adaptation for noncommercial purposes with appropriate attribution, a link to the licence, an indication of changes, and attribution that preserves the licensor's independence from the reuse. Reuse is governed solely by that licence; the responsible-use request on the preceding page remains separate from its terms. The licence deed and legal-code link are available at `https://creativecommons.org/licenses/by-nc/4.0/`. The licence governs in case of conflict with this summary. Third-party material remains subject to the rights held by its respective rights holders.

Statement on the use of language models. The exploratory discussions and preparation of this paper involved OpenAI's ChatGPT. ChatGPT supported exploratory dialogue, conceptual reconstruction, source discovery followed by verification, argumentative criticism, and drafting in LaTeX. The author selected the research questions, directed and approved the theoretical commitments and epistemic status of the claims, and bears sole responsibility for the manuscript, including its definitions, formal constructions, arguments, conclusions, and errors. Authorship credit remains with the human author.

References

[1] Soheil Abbasloo. "Measuring Reasoning in LLMs: A New Dialectical Angle". In: *arXiv preprint arXiv:2510.18134* (2025).

[2] Eser Aygün et al. "An AI System to Help Scientists Write Expert-Level Empirical Software". In: *Nature* 654 (2026), pp. 909–916. DOI: 10.1038/s41586-026-10658-6.

[3] Alexander Bird. "Thomas Kuhn". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2025. Accessed 16 September 2026. Metaphysics Research Lab, Stanford University, 2025.

[4] Daniil A. Boiko et al. "Autonomous Chemical Research with Large Language Models". In: *Nature* 624 (2023), pp. 570–578. DOI: 10.1038/s41586-023-06792-0.

[5] Yilun Du et al. "Improving Factuality and Reasoning in Language Models through Multiagent Debate". In: *arXiv preprint arXiv:2305.14325* (2023).

[6] Ali E. Ghareeb et al. "A Multi-Agent System for Automating Scientific Discovery". In: *Nature* 655 (2026), pp. 497–505. DOI: 10.1038/s41586-026-10652-y.

DRAFT

[7] Google DeepMind. *AlphaEvolve: A Gemini-Powered Coding Agent for Designing Advanced Algorithms*. Google DeepMind Blog, 14 May 2025. Accessed 16 September 2026. 2025.

[8] Google Research. *Accelerating Scientific Breakthroughs with an AI Co-Scientist*. Google Research Blog. Accessed 16 September 2026. 2025.

[9] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. "Towards an AI Co-Scientist". In: *arXiv preprint arXiv:2502.18864* (2025).

[10] Phillip Isola et al. *Digital Red Queen: Adversarial Program Evolution in Core War with LLMs*. Sakana AI research report, 8 January 2026. Accessed 16 September 2026. 2026.

[11] David Leopold. "Analytical Marxism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2022. Accessed 16 September 2026. Metaphysics Research Lab, Stanford University, 2022.

[12] Jiaqi Liu, Shi Qiu, Mairui Li, et al. "AutoResearchClaw: Self-Reinforcing Autonomous Research with Human-AI Collaboration". In: *arXiv preprint arXiv:2605.20025* (2026).

[13] Yexiang Liu et al. "Breaking Mental Set to Improve Reasoning through Diverse Multi-Agent Debate". In: *International Conference on Learning Representations*. 2025.

[14] Helen Longino and Kristina Rolin. "The Social Dimensions of Scientific Knowledge". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2026. Current version consulted 16 September 2026. Metaphysics Research Lab, Stanford University, 2026.

[15] Chris Lu et al. "Towards End-to-End Automation of AI Research". In: *Nature* 651 (2026), pp. 914–919. DOI: 10.1038/s41586-026-10265-5.

[16] Karl Marx. *Preface to A Contribution to the Critique of Political Economy*. Marxists Internet Archive. Primary text; accessed 16 September 2026. 1859.

[17] Julie E. Maybee. *Hegel's Dialectics*. The Stanford Encyclopedia of Philosophy. Substantive revision 2 October 2020; accessed 16 September 2026. 2016.

[18] Thomas Nickles. "Scientific Revolutions". In: *The Stanford Encyclopedia of Philosophy*. Substantive revision 28 November 2017; accessed 16 September 2026. Metaphysics Research Lab, Stanford University, 2017.

[19] Bernardino Romera-Paredes et al. "Mathematical Discoveries from Program Search with Large Language Models". In: *Nature* 625 (2024), pp. 468–475. DOI: 10.1038/s41586-023-06924-6.

[20] Rui Wang et al. "Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions". In: *arXiv preprint arXiv:1901.01753* (2019).

[21] Jiaqi Wei, Yuejin Yang, Xiang Zhang, et al. "From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery". In: *arXiv preprint arXiv:2508.14111* (2025).

[22] Haolun Wu, Zhenkun Li, and Lingyao Li. "Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning". In: *arXiv preprint arXiv:2511.07784* (2025).

[23] Hongliang Xin, John R. Kitchin, and Heather J. Kulik. "Towards Agentic Science for Advancing Scientific Discovery". In: *Nature Machine Intelligence* 7 (2025), pp. 1373–1375. DOI: 10.1038/s42256-025-01110-x.

[24] Yutaro Yamada et al. "The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search". In: *arXiv preprint arXiv:2504.08066* (2025).

[25] Xunjian Yin et al. "Gödel Agent: A Self-Referential Agent Framework for Recursive Self-Improvement". In: *arXiv preprint arXiv:2410.04444* (2024).

[26] Jenny Zhang et al. "Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents". In: *arXiv preprint arXiv:2505.22954* (2025).

DRAFT