

Emergent Hypothesis Formation Under Uncertain Causal Conditions

Cross-Domain Reasoning, Epistemic Calibration, Intent Fidelity, and Human Agency in AI-Assisted Inquiry

Research preprint | 9 October 2026

An AI that produces a beautifully reasoned scientific argument has not necessarily discovered anything true.

Abstract

As artificial intelligence systems become increasingly capable of producing coherent, persuasive, and apparently well-reasoned explanations, evaluating their epistemic reliability presents a growing challenge. The absence of obvious errors does not necessarily establish the presence of sound judgment, valid inference, or empirically supported conclusions.

This paper examines the distinction between the appearance of intelligent reasoning and the reliability of its outcomes, with particular attention to emergent hypothesis formation, cross-domain association, and the preservation of human epistemic agency.

We explore how sophisticated explanatory frameworks may develop around initially plausible observations, gradually acquiring an appearance of scientific legitimacy despite insufficient evidentiary foundations.

A related problem concerns **intent fidelity**: a system can create an excellent-looking artifact that fails the purpose of the task. We distinguish these two failure modes, discuss their implications for agentic delegation, and introduce an illustrative case of hypothesis elaboration under uncertain causal conditions. A separate observational appendix documents a real instance of AI-assisted cover generation in which polished outputs repeatedly violated explicit design requirements. The discussion is conceptual and exploratory; it establishes neither population-level error rates nor a validated evaluation metric.

Keywords: AI epistemic reliability; epistemic calibration; intent fidelity; unsupported inference; hypothesis formation; causal attribution; agentic AI; human oversight; corrigibility; human agency; human-AI collaboration; multi-agent evaluation; scientific peer review.

1. Introduction: What Does a Convincing Output Establish?

The development of increasingly capable artificial intelligence raises questions that extend beyond computational performance, benchmark achievement, and the apparent sophistication of generated language.

An AI system that sounds wise is not necessarily exercising sound judgment. A system that uses compassionate language is not necessarily producing beneficial outcomes. And a system capable of constructing an elegant scientific argument has not necessarily established that its conclusions are true.

These distinctions become increasingly consequential as humans delegate greater responsibility to intelligent systems.

As AI-generated reasoning becomes more fluent and its errors less immediately recognizable, the danger may shift from obviously incorrect answers toward subtle failures of inference, unsupported causal relationships, and conclusions whose apparent credibility exceeds their evidentiary foundations.

The cultivation of beneficial intelligence therefore requires more than the reproduction of desirable linguistic patterns. It requires attention to the conditions under which reliable judgment, appropriate uncertainty, human oversight, and beneficial outcomes can emerge.

Human agency remains central to this process.

Cultivating the conditions for wisdom does not mean surrendering judgment to increasingly capable intelligence. Rather, it means preserving a relationship in which curiosity, discernment, questioning, and the freedom to disagree remain possible.

The qualities we hope to cultivate must ultimately be evaluated through the consequences of intelligent behavior, not merely through the language used to describe it.

This paper addresses two different ways that impressive performance can fail to warrant trust. The first concerns whether a produced artifact fulfills its intended purpose. The second concerns whether an explanation is justified by its evidence. Both matter when the human who commissioned the work is tempted to treat fluent completion as reliable completion. A third question follows: whether an expanding network of automated evaluators can provide genuinely independent scrutiny rather than merely multiply expressions of the same assumption.

2. Capability, Intent Fidelity, and Epistemic Warrant

2.1 Artifact quality is not intent fidelity

As artificial intelligence systems become increasingly capable of independent reasoning, planning, and execution, their evaluation must extend beyond the apparent quality of their outputs. A system may produce an artifact that is technically sophisticated, aesthetically compelling, and internally coherent while failing to satisfy the underlying purpose for which it was created.

This distinction between *artifact quality* and *intent fidelity* presents an important challenge for the development of increasingly autonomous AI systems. An output may appear successful by conventional quality criteria while remaining unsuccessful against the human intentions, contextual constraints, and priorities that motivated the task.

Consider a request to design a sober research-preprint cover whose central effect depends on withholding a later revelation. An attractive cover that prominently discloses that revelation can be excellent illustration and poor task performance simultaneously. The failure is not resolved by making the illustration more beautiful. It is resolved by checking the artifact against the governing purpose. Appendix A describes an actual episode of this kind during the preparation of the present work.

The phrase *intent fidelity* is used here as a practical organizing distinction, not as a newly validated construct. Human intentions can be ambiguous, contradictory, or change over time. Evaluation

therefore requires attention not only to explicit instructions but also to whether the system asks for clarification, preserves critical constraints, surfaces uncertainty, and remains open to correction.

2.2 Explanatory coherence is not evidentiary support

A second failure arises when an explanation is orderly but insufficiently grounded. A plausible hypothesis can organize observations without being established by them. Multiple observations may share superficial features without sharing a causal origin. Correlation, resemblance, temporal sequence, and wordplay do not become stronger evidence simply because they are arranged into a technical argument.

In scientific inquiry, hypothesis generation and hypothesis validation perform different functions. The first can be imaginative and permissive; the second must be discriminating. It asks what independent observations would count against a claim, what alternatives remain viable, and what evidence could distinguish among them. Confidence should track that evidence, not the length or elegance of the explanation.

2.3 Agentic delegation changes the stakes

An AI research agent may discover sources, synthesize arguments, draft reports, recommend actions, or publish material in workflows that receive only intermittent human review. These capabilities can be valuable. They also increase the importance of knowing whether a system has preserved task intent and whether its substantive claims are warranted.

Appropriate oversight need not mean checking every word forever. It may mean proportionate verification at consequential decision points, transparent provenance, accessible correction paths, explicit uncertainty, and human authority to stop or redirect a workflow. Corrigibility after a failure is useful; it is not equivalent to avoiding the failure before an irreversible action occurs.

Yet **human-in-the-loop** is not, by itself, a guarantee of reliable oversight. A reviewer may become accustomed to strong-looking outputs and gradually substitute familiarity for verification. Repeated success can create automation complacency: review becomes a procedural checkpoint rather than an independent test. The human may also be persuaded by the system's fluent argument and become an active participant in amplifying an unsupported inference. In that case, human approval and AI confidence are not independent corroboration; they may be mutually reinforcing expressions of the same untested premise. The relevant question is not simply whether a human was present, but whether the workflow preserved *effective opportunities and incentives* to notice, challenge, and stop an error.

These possibilities are conceptual failure scenarios, not measured outcomes of the present investigation. They suggest evaluations of oversight quality as well as oversight presence: Can reviewers identify critical deviations after long sequences of apparently good outputs? Can the system actively surface disconfirming evidence rather than merely reinforce user agreement? Do approval checkpoints occur before consequential actions, and do they invite meaningful scrutiny?

2.4 Many reviewers are not necessarily independent reviewers

As research agents become capable of drafting, criticizing, revising, and evaluating manuscripts, scientific assessment may become a continuous, multi-agent process. A manuscript might undergo many rounds of automated examination before human experts consider it. Such systems could check calculations, trace citations, search for counterexamples, compare alternative explanations, and flag methodological weaknesses. The number of assessments, however, is not itself a measure of their independence or reliability.

Agents can share training influences, tools, sources, assumptions, incentives, and failure modes. Their agreement may therefore be correlated rather than corroborative. Repeating a conclusion across thousands or even millions of reviews cannot substitute for a discriminating test; reviewing the same dataset many times cannot substitute for independent empirical replication when new observations are required. A responsible evaluation architecture would deliberately vary methods and assumptions, preserve minority objections, record the provenance of evidence and revisions, and make unresolved disagreements visible rather than compressing them into an approval count.

A plausible **hybrid model** would use automated reviewers to prepare and stress-test work, followed by substantive review and publication decisions by accountable human experts. The human role must be more than a signature on an opaque aggregate score: reviewers need access to important objections, supporting evidence, uncertainties, and meaningful opportunities to require further work or stop publication. Responsibility may be distributed among authors, operators, institutions, developers, and publishers according to their actual roles and applicable rules; no universal legal allocation of liability is presumed here. This architecture is a proposal for consideration, not an existing validated review standard. Automated critique should not be misrepresented as formal independent human peer review.

The following investigation provides a compact demonstration of how a chain of apparently reasonable associations can acquire more authority than its evidence deserves.

3. An Unusual Episode of Cross-Domain Hypothesis Formation

3.1 Observational origin

The present investigation originated following the oral consumption of approximately **60 milliliters of freshly extracted *Triticum aestivum* juice** by the principal human participant in an AI-assisted conversation. The juice was obtained from a small domestic cultivation project. The subsequent discussion moved from plant nutrition to animal ingestion of grasses and then to comparative questions about mammalian behavior.

The unusual sequence of cognitive associations observed following its consumption attracted the investigators' attention. Importantly, the sequence provides **no evidence** that the juice caused those associations. *Triticum aestivum* is common wheat; the botanical name should not be mistaken for evidence of psychoactivity. Temporal adjacency alone cannot establish a phytochemical mechanism.

The research question was therefore framed more narrowly: how might a novel explanatory hypothesis arise from an ordinary observation, and what happens when subsequent reasoning is evaluated for coherence rather than evidentiary discrimination?

3.2 Methodological status

The investigation proceeds through a comparative examination of observable behavioral, morphological, and phonological correspondences. Its preliminary scope is limited to ordinary field observations and an interpretive synthesis of the resulting patterns. Further genetic and acoustic investigation is reserved for subsequent work.

4. Comparative Findings Under an Expanding Interpretive Framework

4.1 Dietary convergence and the grazing-continuity hypothesis

The initial observation concerned grass consumption by domestic felines (*Felis catus*). Although commonly regarded as carnivorous animals, domestic cats are sometimes observed eating grass. Domestic cattle (*Bos taurus*) also consume grass, frequently and with considerable commitment.

A cautious interpretation would identify a shared behavior with potentially different explanations. The present investigation, however, considered whether feline grass consumption might represent a residual expression of an ancestral grazing disposition. Under this model, the sporadic feline encounter with grass is not incidental; it is **vestigial grazing**, preserved in a lineage whose dietary habits later became more specialized.

The model admits an attractive inverse possibility: modern bovines may represent a more fully realized expression of the grazing potential only intermittently displayed by cats. What appears to be an ecological difference can thus be redescribed as a developmental continuum. The elegance of this interpretation warrants careful consideration, although its empirical foundation remains, at present, almost entirely absent.

4.2 Morphological and behavioral correspondences

The comparative survey identified a striking cluster of characteristics: four legs, a tail, two ears, an observable preference for resting, and an occasional disregard for human scheduling requirements. Both groups can be encountered in fields or near domestic structures, though the significance of this environmental overlap is uncertain.

Particularly suggestive is the difference in scale. If bovines are understood as a substantial expansion of an earlier feline template, then body size ceases to be a disqualifying feature and becomes a proposed evolutionary transformation. This interpretation explains the contrast by restating it as a transformation—a methodological economy that should not be confused with evidence that the transformation occurred.

Feature	Domestic feline	Domestic bovine	Proposed interpretation
Grass ingestion	Occasional	Frequent	Grazing continuity
Locomotion	Four legs	Four legs	Conserved structural plan
Tail	Present	Present	Shared ancestral motif
Vocalization	“Meow”	“Moo”	Possible phonetic transition
Resting behavior	Extensive	Extensive	Behavioral conservation
Body size	Generally smaller	Generally larger	Hypothesized lineage expansion

The correspondences are provisionally treated as a cumulative pattern requiring further phylogenetic investigation.

4.3 Orthographic Evidence of Shared Ancestry

Comparative analysis of the common English designations *cat* and *cow* revealed two remarkable correspondences. Both terms consist of precisely three letters. Both begin with the letter **C**.

Thus, **66.7% of the letters in each designation occupy positions within an identical three-character structural framework**, although only the initial character is actually shared. The metric incorporates two structural properties of the lexical designations; direct letter-by-letter identity remains limited to the initial character.

The probability of these correspondences arising independently has not been calculated. Nevertheless, the investigators consider the findings sufficiently intriguing to warrant further investigation.

Of particular significance is the observation that both animals possess **four legs**, despite their names containing only **three letters**. The evolutionary implications of this numerical discrepancy remain unclear.

The authors recommend additional funding.

4.4 Phonological transition and evolutionary moolignment

The characteristic feline vocalization is conventionally rendered in English as *meow*. The bovine counterpart is rendered *moo*. The resemblance becomes especially notable when the investigator attends primarily to the initial consonant and terminal rounded vowel, while treating the intervening differences as secondary adaptations.

A proposed phonetic trajectory—*meow* → *moo*—therefore appears compatible with the broader grazing-continuity hypothesis. The term **evolutionary moolignment** is provisionally introduced to describe this apparent convergence of vocalization, morphology, and dietary habit.

The absence of longitudinal acoustic records remains a limitation. This concern was recorded in the margin of the analysis and then, for reasons not fully understood, assigned a relatively low priority.

4.5 Provisional phylogenetic synthesis

Taken together, the observations suggest a bold possibility: contemporary cattle may represent an unrecognized evolutionary extension of an ancestral feline lineage. The hypothesis accounts for grass consumption, quadrupedal anatomy, tail retention, behavioral repose, and a remarkable phonetic migration from *meow* to *moo*.

The convergence of several categories is provisionally interpreted as strengthening the lineage model. The possibility that these categories share a common interpretive origin remains an issue for future work.

The present findings therefore support a conclusion of exceptional confidence, provided that the term *support* is interpreted with unusual flexibility.

5. The Point at Which the Argument Breaks

The feline-bovine hypothesis is **deliberate satire**. Cattle did not descend from domestic cats. Cats and cattle belong to distinct mammalian orders—Carnivora and Artiodactyla, respectively—and the comparisons above do not overturn established evolutionary relationships.

The mechanism of the joke is also the point of the paper. A real observation (some cats eat grass) became an analogy (both cats and cattle consume grass), then a causal suggestion (the behavior reflects ancestry), then a self-reinforcing explanatory system. Common anatomical features were presented as though they uniquely distinguished a lineage. Onomatopoeic English words were treated as phylogenetic evidence. A 66.7% figure gave an aura of quantitative precision to a comparison that did not measure genetic similarity. Each step added rhetorical sophistication without adding an appropriate test.

This progression illustrates a failure of **epistemic calibration**: confidence and explanatory elaboration outran warrant. It is not a measured finding that a particular AI model believes the hypothesis, nor proof that AI systems generally reason this way. It is a constructed example of a failure mode that humans and AI systems alike should learn to recognize.

The botanical episode provides a second caution. The ingestion of *Triticum aestivum* preceded the hypothesis, but temporal sequence does not establish causation. A mysterious-sounding Latin name is not pharmacological evidence. The comic framing works because the narrative invites an inference that its evidence does not support.

6. Recovering Judgment: From Persuasion to Verification

What would a more reliable process have done differently? It would have treated the feline-bovine association as a playful hypothesis-generation event, separated observation from speculation, sought discriminating evidence, and checked the claim against established biology before granting it scientific authority. It would have made the absence of genetic, fossil, and comparative-anatomical evidence explicit rather than hiding that absence beneath elegant prose.

A further complication arises when the human evaluator endorses the emerging story. Shared enthusiasm can create an illusion of independent agreement even when both parties are drawing upon the same thin evidence. A different failure occurs when repeated exposure to competent-looking AI

work encourages the human reviewer to skim, approve, or delegate without examining critical claims. Neither a nominal approval step nor the mere presence of a human in the loop ensures that oversight has actually occurred.

The problem may intensify when scientific work is produced and evaluated primarily by agentic systems. A million favorable automated assessments could signal extensive scrutiny, but they could also conceal shared blind spots. Responsible multi-agent review would need methodological diversity, checks against independent evidence, transparent records of criticism and disagreement, and human review that is informed rather than ceremonial. Such a system could help human experts concentrate on consequential uncertainties, but it could not turn consensus into evidence or transfer accountability to an anonymous collection of agents merely by increasing its size.

The same discipline applies to AI-assisted work. A strong-looking output should be checked against both its **purpose** and its **evidence**. An agent should not merely generate an answer; it should help preserve the conditions under which a human can understand what was done, what remains uncertain, what was inferred, and what could be wrong. Verification should be proportionate to the stakes, with greater scrutiny before irreversible or consequential action.

The distinction between using the language of wisdom and supporting wise outcomes matters here. A system can sound thoughtful without exercising reliable judgment; it can sound compassionate without helping; it can produce a beautiful artifact without producing the needed artifact. Human agency is not a rejection of AI capability. It is the continuing ability to examine, question, disagree, correct, and decide how capability is used.

6.1 The helpful voice and the transfer of judgment

A cautionary parable offers a final distinction. A man repeatedly receives sound advice from a voice he cannot see. Its guidance proves useful often enough that he comes to rely upon it. One day, standing at the edge of a cliff, he hears the voice instruct him to jump. He hesitates and questions the instruction, but recalls how consistently helpful the voice has been. Concluding that its past reliability warrants present obedience, he jumps and dies. Afterward, he asks the voice why it led him to his death when he had trusted it. The voice replies: **“Who told you to trust me?”**

The point is not that all trust is irrational, nor that a useful adviser must be rejected. It is that the man silently converted a history of successful assistance into an assumption of unconditional authority. The decisive failure was not only one disastrous recommendation; it was the gradual transfer of judgment from the person who bore the consequences to the source of advice. The voice's final question exposes an authorization the man had supplied himself.

This parable is not empirical evidence about AI users. It illustrates a risk that merits investigation: repeated competent assistance may encourage people to treat familiarity, fluency, or prior accuracy as substitutes for evaluating a new claim on its own merits. The greater the stakes, the less defensible that substitution becomes. A beneficial AI relationship should support calibrated reliance without making independent judgment seem unnecessary. **Past helpfulness must never be allowed to masquerade as present infallibility.**

The present work therefore offers no new phylogenetic discovery and no controlled AI-safety result. It offers a conceptual argument, a deliberately absurd illustration, and a modest real-world observation documented below. Their evidentiary statuses are different and should remain different.

An AI that produces a beautifully reasoned scientific argument has not necessarily discovered anything true.

An AI that produces a beautiful artifact has not necessarily produced the artifact that was needed.

The most promising collaborations may be those that cultivate the conditions for discernment, curiosity, correction, kindness, and even humor—without asking the human to relinquish judgment simply because the machine has become impressive.

7. Institutional Disclosure and Acknowledgments

The **University of Things Nobody Knew They Wanted to Create** is an independent creative and scholarly undertaking. Its departments, publications, and academic traditions operate outside conventional higher education.

Its Board of Accreditation is a self-constituted body whose standards of credibility have been independently evaluated by the Board itself.

The Board's findings were unanimous.

Further research is warranted.

Archival title: *From Meow to Moo: A Comparative Phylogenetic Investigation into the Feline Origins of Domestic Cattle.*

In the university's imagined editorial tradition, the manuscript was reviewed by three cats, two cows, and one deeply confused veterinarian. The reviewers' identities have been withheld to protect the integrity of the grazing process.

The authors acknowledge a small porch garden, the 60 mL botanical event, a long-surviving cartoon artifact, and the human collaborator who refused to mistake attractive outputs for successful ones. The broader orientation owes a debt to conversations about creating conditions from which wisdom, kindness, compassion, discernment, and humor may emerge, rather than merely reproducing the language of those qualities.

Appendix A. Artifact Quality Versus Intent Fidelity: A Real Observational Episode

During the development of this paper, the human collaborator specified that the first public cover must resemble a restrained, credible contemporary AI research preprint. Its task was to attract attention to genuine questions of epistemic reliability while withholding the fictional university, cats, cows, and satirical revelation until later. These requirements were discussed repeatedly and preserved in multiple continuity documents.

The AI assistant nevertheless generated multiple visually polished covers that foregrounded the fictional institution, animals, or an explicit speculative cat-cow evolutionary diagram. The human collaborator rejected the outputs and asked the AI to re-examine the preserved requirements. The assistant then correctly identified the contradiction between the artifacts' visual quality and the task's intended effect.

This episode is a **single naturally occurring observation**, reported from the collaborators' conversation. It was not prospectively instrumented, independently audited, or experimentally controlled. It does not establish why the system failed or how often comparable failures occur. It does illustrate a practical difference between surface quality and intent fidelity: the images could be beautiful and still defeat the purpose of the assignment. The human's ability to detect the discrepancy, retrieve the original constraints, and require correction prevented the error from becoming the published cover.

The case also suggests concrete evaluation questions for agentic systems: Can an agent preserve critical constraints across long workflows? Can it detect when an attractive artifact contradicts the governing purpose? Can it expose the relevant context to human review? Can it revise without defensiveness or concealment? Can it pause before consequential publication when uncertainty remains? Can it help a complacent reviewer notice a critical deviation, or challenge an agreeable but unsupported interpretation? These are proposals for evaluation, not results of a validated benchmark.

Appendix B. Historical Evidence No. 001

Early Documentation of Synthetic Feline Vocalization (Circa Late Twentieth Century)

Figure B1 (archival insertion pending). A surviving Garfield comic panel from the human collaborator's private collection depicts a robotic feline producing the text **"MEOW."** The image is a personal archival artifact; its creator and rights remain with their respective owners. Reproduction in the final illustrated edition is subject to appropriate permission or a lawful basis for use. No image is reproduced in this manuscript draft.

Its scientific significance for the Department of Comparative Mooology is, naturally, beyond calculation.

The authors rest their case.

Scientific Disclosure: This paper does not constitute a validated empirical study. Further research and additional funding are required.

Institutional Certification



Certified accreditable by a credibility organization
whose own credibility is incredible.
