The Principle of Obviousness: Representation, Compression, and the Erasure of Insight

Anonymous

December 11, 2025

Abstract

We draw a sharp distinction between three notions that human agents routinely conflate: (i) *obviousness*—a result is easy to see in a particular representation, (ii) *triviality*—a result would have been easy to see in any reasonable representation already available, and (iii) *truth*—a result is in fact reliably grounded in the world. We formalise *obviousness* and *triviality* in terms of representation-dependent cognitive cost, and we show how a shift to a more compressed representation can make a hard-won result feel retrospectively trivial. We call this systematic misclassification *McCaul's Principle of Obviousness*. The principle explains a common but under-analysed phenomenon: once a powerful new representation has been learned, agents downgrade the insight it provides, insisting that they "knew it all along" and erasing the representational work that made the result cheap. We present a simple formal model of representation change and memory distortion, derive behavioural predictions, and sketch experimental paradigms to test them. Finally, we connect the principle to hallucinations in large language models: these systems are explicitly optimised to make continuations obvious in their internal code, while neither the model nor the human user reliably pays the cost of checking truth. Obviousness, we argue, is evidence about compression, not about triviality or truth; failing to track this distinction has consequences for scientific credit, pedagogy, and the evaluation of human–AI collaborations.

1 Introduction

Some of the most powerful ideas in science and mathematics share a peculiar fate. Once the right diagram, formalism or change of variables has been introduced, a previously opaque result becomes "obvious". At that moment, non-innovators frequently react not with gratitude but with downgrading: *"Well, that's trivial; I basically knew that."* The very success of the explanation undermines recognition of the insight.

This paper makes three simple claims.

First, we argue that obviousness is a *property of a representation*, not of the underlying content. A result may be hard to see in one representation and effortless in another; the difference is the cost of the representational mapping.

Second, we formalise a principle—*McCaul's Principle of Obviousness*—which states that once a representation has made a result obvious, agents tend to misremember its prior

difficulty and misclassify the result as trivial. The cognitive error is not just hindsight bias in the usual sense; it is a specific failure to track the cost of building the representation that made the result cheap.

Third, we show how this principle interacts with modern AI systems. Large language models (LLMs) and related architectures are explicitly optimised to learn internal codes in which complex continuations become easy to predict. They are, in a precise sense, engines for making things *obvious in representation space*, with no direct term for truth. Hallucinations can be seen as continuations that are maximally obvious in the model's code but unverified in the world. When humans then read these fluent outputs, they are subject to McCaul's principle: they experience the answers as obvious, retroactively treat them as trivial, and are tempted to assume they are therefore true.

We develop these ideas as follows. Section 2 defines obviousness and triviality in terms of representation-dependent cognitive cost. Section 3 states McCaul's Principle of Obviousness and explores its conceptual consequences. Section 4 presents a simple formal model in which representation change induces systematic misestimation of past difficulty. Section 5 connects the framework to hallucinations and LLM behaviour. Section 6 sketches empirical paradigms to test the principle. Section 7 discusses implications for science, pedagogy, and AI interpretability.

2 Obviousness, triviality, and cost

We begin by putting names to three notions that are often conflated in informal talk.

2.1 Representations and cognitive cost

Let $\mathcal{W}$ be some domain of interest: the world, a mathematical structure, or a task environment. A *representation* R is, abstractly, a code or model that supports reasoning about $\mathcal{W}$: for example, a choice of coordinate system, a diagrammatic language, a statistical model, or a programming language. Different representations can make different aspects of $\mathcal{W}$ easy or hard to see.

Let P be a proposition or pattern about $\mathcal{W}$ (e.g. a theorem, a causal relationship, or a strategy). We introduce a coarse-grained notion of *cognitive cost*:

$$\mathcal{C}_R(P) \in [0, \infty) \tag{1}$$

denotes the minimal cost, for a given agent, of deriving, recognising or reliably working with P *within* representation R. The cost may be measured in steps of reasoning, time, description length, or any other reasonable proxy; for present purposes, we treat it as an abstract scalar.

The key point is that $\mathcal{C}_R(P)$ is representation-dependent: switching to a different R' can dramatically change the cost of the same proposition.

2.2 Obviousness in a representation

We can now define:

Definition 1 (Obviousness). We say that P is *obvious in representation R for an agent* if, once the agent has internalised R, the cost $\mathcal{C}_R(P)$ is small relative to the agent's typical reasoning budget. Informally: from within R, P is "easy to see."

This notion is subjective and relative: it depends on both R and the agent. It is also *time-asymmetric*: a result can move from non-obvious to obvious as the agent becomes familiar with R.

2.3 Triviality as representation-invariance

By contrast, we use *trivial* to describe propositions that were low-cost all along.

Definition 2 (Triviality). We say that P is *trivial* for an agent if $\mathcal{C}_R(P)$ is small for all reasonable representations R that were available to that agent prior to some new representational insight. Informally: the agent could have seen P cheaply *without* any substantive change of frame.

Triviality is thus a stronger, representation-invariant claim. Obviousness is about current cost in a particular R; triviality is about historical cost across the agent's prior representational repertoire.

2.4 Truth as a separate axis

Obviousness and triviality are purely internal properties: they concern the relationship between P and the agent's codes. *Truth*, by contrast, is about the relationship between P and $\mathcal{W}$. A proposition may be obvious and false, trivial and false, obvious and true, or non-obvious and true. Our concern in this paper is with the systematic conflation of these axes.

3 McCaul's Principle of Obviousness

We can now state the central principle.

3.1 Informal statement

In everyday terms:

McCaul's Principle of Obviousness (informal). *Once a representation has compressed a problem so that some result looks obvious, observers will systematically misclassify that result as trivial, and will often insist they "knew it all along," thereby erasing the cost of the compression that made it obvious.*

The key claim is not just that people are overconfident after learning something. It is that they specifically fail to track the *representational work* that was done: the cost of moving from a high-cost $\mathcal{C}_{R_0}(P)$ in an old representation to a low-cost $\mathcal{C}_{R_1}(P)$ in a new one.

3.2 Formal statement

To make this precise, consider an agent with an initial representation R_0 and a proposition P such that

$$\mathcal{C}_{R_0}(P) \gg 1. \tag{2}$$

Suppose the agent subsequently acquires a new representation R_1 such that

$$\mathcal{C}_{R_1}(P) \ll \mathcal{C}_{R_0}(P). \tag{3}$$

That is, R_1 makes P cheap to see.

We can define the *compression gain* for P as

$$\Delta(P; R_0 \to R_1) := \mathcal{C}_{R_0}(P) - \mathcal{C}_{R_1}(P). \tag{4}$$

In an ideal agent with perfect meta-cognition and memory, this gain would be recognised: the agent would remember that P was hard to see in R_0, and would attribute credit for the reduction in cost to the new representation.

McCaul's principle is that real agents do not behave this way.

McCaul's Principle of Obviousness (formal). Let $R_0 \to R_1$ be a representational change for an agent, and let P be such that $\Delta(P; R_0 \to R_1) \gg 0$. Then, after internalising R_1, the agent will tend, by a combination of hindsight bias and fluency illusion, to:

1. Underestimate $\mathcal{C}_{R_0}(P)$ in retrospect (i.e. misremember P as having been easier to see in R_0 than it was), and

2. Reclassify P as trivial, i.e. behave as if $\mathcal{C}_R(P)$ had been small across all prior representations, rather than attributing the cost reduction to the specific change $R_0 \to R_1$.

Equivalently: the agent will collapse the distinction between "obvious in R_1" and "trivial all along," erasing the representational contribution.

3.3 Conceptual consequences

Several consequences follow.

First, the better an explanation is—in the sense of making a wide class of results obvious in a new code—the more vulnerable it is to being dismissed as trivial once learned. This is not a paradox; it is the natural consequence of forgetting the initial cost landscape.

Second, the phenomenon is asymmetric in status. Non-innovators can defend their self-image by reclassifying the innovator's contribution as trivial, rather than admitting that they lacked the representation which made P cheap.

Third, the principle suggests that a large fraction of scientific and mathematical credit disputes can be understood as disagreements not about the current cost of P in the accepted representation, but about the historical cost and about who built the representation that made P obvious.

In the remainder of the paper we show that the principle can be embedded in a simple formal model and that it yields testable behavioural predictions.

4 A simple model of representation change and misremembered cost

We outline a minimal model to make McCaul's principle more precise. The goal is not to capture the full complexity of human cognition, but to show that under plausible assumptions, a representation change $R_0 \to R_1$ generically leads to underestimation of past difficulty.

4.1 Setup

Consider an agent confronted with a family of propositions $\{P_i\}$. For each P_i, in each representation R, there is a true underlying cost $\mathcal{C}_R(P_i)$. We suppose that in R_0, some subset S of propositions are high-cost:

$$\mathcal{C}_{R_0}(P_i) = C_i^{(0)} \gg 1, \quad i \in S. \tag{5}$$

The agent initially works in R_0, attempts to derive or understand some P_i, and forms an internal estimate $\hat{C}_i^{\text{pre}}$ of how hard it is. We can model this as noisy observation:

$$\hat{C}_i^{\text{pre}} = C_i^{(0)} + \epsilon_i^{\text{pre}}, \tag{6}$$

with ϵ_i^{pre} mean-zero or with some bias.

Now the agent learns a new representation R_1 in which those same propositions are cheaper:

$$\mathcal{C}_{R_1}(P_i) = C_i^{(1)} \ll C_i^{(0)}. \tag{7}$$

After training on R_1, the agent experiences P_i as obvious and forms a new estimate $\hat{C}_i^{\text{post}}$ of how hard P_i "was". Crucially, this retrospective estimate is not based on direct access to the historical $C_i^{(0)}$, but on memory and current fluency.

4.2 Retrospective misestimation

We model retrospective difficulty estimates as a weighted combination of: (i) noisy memory of prior effort and (ii) current fluency in R_1:

$$\hat{C}_i^{\text{post}} = \alpha\, \hat{C}_i^{\text{pre}} + (1-\alpha)\, f(C_i^{(1)}) + \epsilon_i^{\text{post}}, \tag{8}$$

with $0 \le \alpha \le 1$, some function f that maps current cost to a difficulty rating, and noise ϵ_i^{post}.

When α is small—the agent puts little weight on stored memory relative to present ease—and $C_i^{(1)} \ll C_i^{(0)}$, we obtain:

$$\mathbb{E}[\hat{C}_i^{\text{post}}] \approx \alpha\, C_i^{(0)} + (1-\alpha) f(C_i^{(1)}) < C_i^{(0)}. \tag{9}$$

Thus, on average, the agent underestimates past difficulty.

If f is itself sublinear or saturating (reflecting the fact that very easy tasks are all experienced as "effortless"), the underestimation can be severe. In the limit $C_i^{(1)} \to 0$ and small α, we have

$$\hat{C}_i^{\text{post}} \approx (1-\alpha) f(0), \tag{10}$$

so that tasks that were originally very hard are now remembered as essentially effortless.

4.3 Reclassification as trivial

Triviality, recall, is the judgement that $\mathcal{C}_R(P)$ was low in *any* prior representation. In practise, agents may not explicitly inspect other representations; instead, they may apply a rule of thumb:

If $\hat{C}_i^{post} \leq \tau$ for some small threshold τ, classify P_i as trivial.

For sufficiently large compression gain and sufficiently small α, many P_i with originally high $C_i^{(0)}$ will cross this threshold. In this way, a representational change induces both underestimation of past difficulty and inappropriate reclassification of non-trivial results as trivial.

This toy model can be elaborated in many ways: we can allow for heterogeneous agents, social transmission of representations, and status dynamics. The basic mechanism, however, is simple: current fluency "leaks backwards" in time and contaminates memory of effort.

5 Obviousness without truth: hallucinations in AI

The distinction between obviousness, triviality and truth becomes particularly acute in the context of large language models and related AI systems.

5.1 LLMs as obviousness engines

At a high level, an LLM is trained to minimise an autoregressive loss: given a context x, predict the next token y. Training adjusts parameters so that the model learns internal representations in which the conditional distribution $p_\theta(y \mid x)$ is as *peaked* as possible on the tokens that occur in the training data.

One can view this as a form of compression: the model discovers internal codes in which likely continuations become *obvious* in the representational sense—they are low-cost to generate and assign high probability.

Note that there is no explicit term in the objective for *truth*. The model is rewarded for matching the distribution of observed text, not for tracking the world. To the extent that training data contain reliable information about the world, truth emerges only indirectly.

5.2 Hallucinations as obvious but untrue

From this perspective, a hallucination is a continuation that is highly probable under $p_\theta(\cdot \mid x)$, and thus maximally obvious in the model's internal representation, but is not in fact grounded in $\mathcal{W}$. Nothing in the training objective prevents such outputs in contexts where the model has not learned a reliable mapping from text to reality.

When a human user reads a hallucinated output, two things happen:

1. The text is often fluent, coherent, and structurally familiar. In the user's own representation—their language and conceptual repertoire—the answer feels obvious.

2. McCaul's principle then applies: the user is tempted to treat the now-obvious answer as trivial, and—critically—to assume that obviousness implies truth.

Thus hallucinations exploit two levels of compression: the model's internal representation, which makes the answer obvious to the machine, and the user's linguistic-cognitive representation, which makes the answer obvious to the human. At neither level is there an automatic guarantee of truth.

5.3 The cost of verification

The missing ingredient is the cost of checking. Determining whether a proposition P is true in $\mathcal{W}$ generally involves a verification procedure with its own cost $V(P)$—running an experiment, consulting a database, or performing a proof. In many real-world settings, $V(P)$ is non-trivial.

When users accept obviousness as a proxy for truth, they implicitly *set* $V(P) \approx 0$ *in their decision policy.* McCaul's principle predicts that once a model has made a particular pattern of answers feel obvious, users will systematically underestimate $V(P)$ and so under-invest in verification.

We can thus summarise:

Hallucinations are answers that are obvious in representation space and untested in reality space. McCaul's principle explains why, once we have seen them, we are tempted to treat them as trivial and therefore true.

This has direct implications for the design of human–AI systems: interfaces and training regimes must actively counter the tendency to equate obviousness with triviality and truth.

6 Empirical predictions and experimental paradigms

We now sketch experiments that could empirically test McCaul's principle.

6.1 Experiment 1: pre/post difficulty ratings

Design. Participants are given a set of problems or propositions $\{P_i\}$ formulated in an initial representation R_0 (for example, algebraic problems in raw symbolic form). For each P_i, participants: (i) attempt to solve the problem, (ii) report whether they succeeded, and (iii) rate its difficulty on a numerical scale; these ratings $\hat{C}_i^{\text{pre}}$ approximate $\mathcal{C}_{R_0}(P_i)$.

Participants are then taught a new representation R_1 designed to compress the problem space (for example, a diagrammatic method or a change of variables), and given training until they can solve related problems easily.

After training, participants are asked to: (i) solve variants of the original problems using R_1 and (ii) retrospectively estimate how difficult the original problems were when first encountered.

Prediction. For problems where $\mathcal{C}_{R_1}(P_i) \ll \mathcal{C}_{R_0}(P_i)$, participants' retrospective difficulty ratings $\hat{C}_i^{\text{post}}$ will be systematically lower than their initial ratings, with a bias that increases with the compression gain. Many originally non-trivial problems will be retrospectively classified as "easy" or "obvious."

6.2 Experiment 2: credit and trivialisation

Design. Two groups of participants are exposed to the same representational trick R_1 that makes a family of results $\{P_i\}$ obvious.

Group A is told a brief story about discovery: they are informed that a particular researcher devised R_1 to solve then-hard problems P_i. Group B is simply taught R_1 as the standard approach, with no narrative about its origin.

Both groups are then asked to evaluate: (i) how obvious each P_i now feels, (ii) how *trivial* they judge P_i to be as a mathematical or conceptual contribution, and (iii) how much credit the inventor of R_1 (if any) deserves.

Prediction. Both groups will report similar levels of current obviousness, but Group B will rate P_i as more trivial and attribute less credit to the originator. This dissociation between obviousness and perceived contribution is expected under McCaul's principle.

6.3 Experiment 3: AI-assisted hallucinations

Design. Participants query an LLM on factual questions where the model is known to sometimes hallucinate (e.g. fabricated references, incorrect but fluent explanations). The interface is modified to collect, for each answer: (i) a rating of how obvious or intuitive the answer feels, (ii) a confidence judgement about its truth, and (iii) a willingness to spend effort verifying it.

Ground truth is available to the experimenter, and verification cost can be manipulated (e.g. by making fact-checking easy or tedious).

Prediction. Answers that are more fluent and feel more obvious will be trusted more and checked less, independent of their actual truth. After being shown corrections to hallucinated answers, participants will tend to underestimate how misleading the original output was and may report that the correction is "obvious in retrospect".

This would support the claim that obviousness in representation space is being inappropriately used as a cue for truth.

7 Discussion

7.1 Scientific credit and the value of representation

McCaul's principle suggests a simple lens on recurrent debates about scientific and mathematical credit. The core insight is that we are poor historians of our own representational progress. Once a field has internalised a powerful formalism, the results it made possible look trivial within that formalism.

This under-recognition of representational work is particularly acute when: (i) the new representation induces large compression gains for many P_i, and (ii) the community rapidly teaches R_1 as the default, without preserving the difficulty of R_0.

Systematically tracking $\mathcal{C}_R(P)$ across representations is not feasible in practise, but awareness of the principle can inform more careful narratives of discovery and more generous attribution of credit to representational innovators.

7.2 Pedagogy and the curse of knowledge

Teachers are notoriously prone to underestimating how opaque material appears to students. McCaul's principle offers a mechanistic account: instructors operate in a representation R_1 where core results are obvious; they have weak access to their own prior state in R_0 and thus systematically misjudge $\mathcal{C}_{R_0}(P)$ for novices.

Pedagogical strategies that explicitly reconstruct the representational journey—showing not just the final code but the path from R_0 to R_1—may mitigate this effect. So might explicit prompts for instructors to estimate not just current obviousness but prior difficulty.

7.3 AI interpretability and epistemic humility

As AI systems increasingly participate in scientific and technical work, the principle of obviousness has two opposing implications.

On the one hand, these systems will help humans learn new representations in which complex patterns become easy to see. The risk is that human users will promptly downgrade these patterns to "trivial," erasing the contribution of both the AI and the representational move.

On the other hand, human users will be exposed to outputs—including hallucinations—that feel obvious but are untested. Without explicit safeguards, they will use obviousness as an implicit proxy for truth, underweighting the cost of verification.

In both cases, the lesson is that obviousness is evidence about compression, not about triviality or truth. Designing interfaces and workflows that make the cost of representation change and verification more salient may be crucial to maintaining epistemic humility.

8 Conclusion

We have argued that a surprisingly pervasive cognitive error can be traced to a simple confusion: humans routinely collapse the distinction between (i) obviousness in a particular representation, (ii) triviality across representations, and (iii) truth about the world. By modelling obviousness and triviality in terms of representation-dependent cost, we formalised *McCaul's Principle of Obviousness*: once a representation has made a result cheap to see, agents tend to misremember its prior difficulty and reclassify it as trivial, thereby erasing the work embodied in the representation.

This principle has concrete implications for scientific credit, pedagogy and the evaluation of AI systems. We sketched simple models and empirical paradigms that can put it to the test. More broadly, the principle serves as a reminder: when a result appears obvious, we should ask not only whether it is true, but also what representational work was done to make it so, and by whom.