The Compressor Gets the Credit

Marvin Vincent Gabler*

Technical Report, April 2026

Abstract

Why do originators of fundamental ideas often fail to receive credit proportional to their causal contribution, while later reformulators who add relatively little novelty capture most of it? We model credit attribution as a consequence of memetic fitness: the rate at which an idea propagates through a network of bounded-compute agents. Using algorithmic information theory, we show that *compression contribution* (reduction in integration cost via reformulation) can dominate *novelty contribution* (reduction in uncertainty about the solution space) in determining adoption, and that this dominance is what bounded-bandwidth populations converge to under standard assumptions. The framework is descriptive: it predicts the allocation that emerges, not the allocation that ought to emerge. We examine three case studies (Lorentz–Poincaré–Einstein, Schmidhuber–Vaswani, LSTM) and show that the allocation that the model predicts is close to the one history records. Under a tractable parameterization the optimal allocation of effort to novelty scales as $1/N$ in a civilization of N agents; the specific exponent depends on scaling choices but the qualitative dominance of compression at scale is robust. We discuss the limitations of the framework and a conditional extension to self-improving systems.

1 Introduction

A recurring pattern in the history of science: Agent A publishes a formal result R at time t_0. Agent B, at time $t_1 > t_0$, publishes a reformulation R' that is mathematically equivalent to R (or a strict superset containing R as a special case) but packaged differently. The community attributes the invention to B. Agent A protests, pointing to priority. The community largely ignores the protest.

This pattern has been documented extensively. Poincaré and Lorentz published the mathematical apparatus of special relativity before Einstein's 1905 paper [Poincaré, 1905, Lorentz, 1904], yet Einstein receives near-exclusive public credit. Schmidhuber published Fast Weight Programmers with outer-product key-value mechanisms in 1991 [Schmidhuber, 1991, 1992], yet Vaswani et al. [2017] receive credit for inventing the Transformer. Konorski published the so-called Hebb rule before Hebb [Konorski, 1948, Hebb, 1949]. Amari published what is called the Hopfield network a decade before Hopfield [Amari, 1972, Hopfield, 1982]. Ivakhnenko published the first working deep learning algorithms in 1965 [Ivakhnenko and Lapa, 1965], decades before the term *deep learning* became fashionable. Linnainmaa published backpropagation in 1970 [Linnainmaa, 1970], years before it was applied to neural networks.

The standard explanation invokes sociology: power dynamics, institutional prestige, the Matthew Effect [Merton, 1968]. While these factors are real, they are not explanatory at a fundamental level. They describe the mechanism but not the optimization target. What is the system actually optimizing for when it allocates credit this way?

We propose that credit attribution in information-processing civilizations, whether human or machine, tracks *memetic fitness*: the rate at which an idea propagates through a population of

*Affiliated with Jua.ai AG, Zurich, Switzerland.

bounded-compute agents. Memetic fitness, in this framework, is dominated by the compression an idea achieves over the receiving agent's prior state rather than by its absolute novelty. The reduction in integration cost that a good reformulation provides is the dominant driver of adoption in our model. The framework is descriptive throughout: it says what propagates, not what ought to be credited. The normative question of whether credit *should* track propagation, or instead counterfactual displacement, is a separate matter we discuss but do not resolve.

2 Formal Framework

2.1 Ideas as Programs

Following Solomonoff [1964] and Schmidhuber's work on algorithmic information theory applied to science [Schmidhuber, 1997, 2009], we model an idea I as a program p on a universal Turing machine U such that $U(p) = I$. The Kolmogorov complexity $K(I)$ of idea I is:

$$K(I) = \min_{p:\, U(p)=I} |p| \tag{1}$$

where $|p|$ denotes the length of program p in bits.

Definition 1 (Idea Space)**.** *Let $\mathcal{I}$ denote the space of all computable ideas (programs that halt on U). An agent α at time t possesses a knowledge state $S_\alpha(t) \subset \mathcal{I}$, representable as a program $s_\alpha(t)$ that enumerates all ideas currently accessible to α.*

Definition 2 (Conditional Complexity)**.** *The complexity of idea I given knowledge state S_α is:*

$$K(I \mid S_\alpha) = \min_{p:\, U(p,s_\alpha)=I} |p| \tag{2}$$

This measures how many additional bits agent α needs to reconstruct I from what it already knows.

Remark 1 (Choice of reference machine)**.** *The definitions above assume a fixed universal Turing machine U. By the invariance theorem, $K_U(x) = K_V(x) + c_{UV}$ for any two UTMs U, V, where c_{UV} is a constant independent of x. In practice, however, agents do not share a reference machine. Each agent's background, training, and notational conventions define an effective UTM U_α, and the true integration cost is $K_{U_\alpha}(I \mid S_\alpha)$, not $K_U(I \mid S_\alpha)$.*

This heterogeneity does not invalidate the framework; it strengthens it. The compression contribution $C(I_1, I_0) = \mathbb{E}_\alpha[K_{U_\alpha}(I_0 \mid S_\alpha) - K_{U_\alpha}(I_1 \mid S_\alpha)]$ averages over the agent-specific UTMs. A reformulation achieves high C precisely when it reduces integration cost across diverse reference machines, that is, when it is expressed in terms that are short programs in many different agent languages simultaneously. This is what distinguishes a good reformulation from a mere re-encoding: Einstein's two-postulate derivation is compact relative to the effective UTMs of physicists, mathematicians, engineers, and educated laypeople alike, whereas Poincaré's ether-dependent formulation is compact only relative to the UTMs of specialists in Lorentzian electrodynamics. The results in this paper hold for any fixed U and, by averaging, for populations of heterogeneous U_α.

2.2 Novelty and Compression Contributions

We now distinguish two types of intellectual contribution.

Definition 3 (Novelty Contribution)**.** *Given a problem domain D with solution space Ω_D, and a community knowledge state $S_C(t)$ at time t, the novelty contribution of idea I published at time t is:*

$$N(I, t) = K(\Omega_D \mid S_C(t)) - K(\Omega_D \mid S_C(t) \cup \{I\}) \tag{3}$$

That is, the reduction in the community's uncertainty about the solution space achieved by I.

Definition 4 (Compression Contribution)**.** *Given an existing idea $I_0 \in S_C(t)$ and a reformulation I_1 such that the denotational content $[\![I_0]\!] = [\![I_1]\!]$ (they solve the same problem or express the same theorem), the compression contribution of I_1 relative to I_0 is:*

$$C(I_1, I_0) = \mathbb{E}_{\alpha \sim \mathcal{A}}[K(I_0 \mid S_\alpha) - K(I_1 \mid S_\alpha)] \tag{4}$$

where $\mathcal{A}$ is the distribution over agents in the receiving community.

The compression contribution measures how much easier I_1 is to integrate, on average, across the community. Note that $C(I_1, I_0) > 0$ is possible even when $N(I_1, t) = 0$, that is, when I_1 adds no new information about the solution space but simply repackages existing information.

2.3 Memetic Fitness

Definition 5 (Memetic Fitness)**.** *The memetic fitness F of idea I at time t in a network of N agents with bounded compute B per unit time is:*

$$F(I, t) = \frac{d}{dt}|\{\alpha \in \mathcal{A} : I \in S_\alpha(t)\}| \tag{5}$$

the rate of adoption of I across the population.

Proposition 1 (Memetic Fitness Depends on Compression)**.** *For agents with compute bound B, the probability that agent α integrates idea I at time t is:*

$$P(\alpha \text{ adopts } I \text{ at } t) \propto \frac{V(I, \alpha)}{K(I \mid S_\alpha(t))} \cdot \mathbf{1}[K(I \mid S_\alpha(t)) \leq B] \tag{6}$$

where $V(I, \alpha)$ is the expected value of I to α, and $\mathbf{1}[\cdot]$ is the indicator function enforcing the compute bound.

Proof. Agent α faces a set of candidate ideas $\{I_1, \ldots, I_m\}$ at time t, each with value $V(I_j, \alpha)$ and integration cost $K(I_j \mid S_\alpha)$. The agent has compute budget B and must choose which ideas to integrate. This is a 0-1 knapsack problem:

$$\max \sum_j x_j V(I_j, \alpha) \quad \text{s.t.} \quad \sum_j x_j K(I_j \mid S_\alpha) \leq B, \quad x_j \in \{0, 1\} \tag{7}$$

The linear programming relaxation ($x_j \in [0, 1]$) is solved by the greedy algorithm: sort ideas by value-to-cost ratio V/K and fill the budget in descending order. In the fractional solution, items are selected with probability proportional to their V/K ratio, subject to the budget constraint. Items with $K(I_j \mid S_\alpha) > B$ are infeasible regardless of value. For a population of agents encountering ideas stochastically, this yields the stated adoption probability in expectation. □

3 Main Results

3.1 The Compression Dominance Theorem

Theorem 2 (Compression Dominance)**.** *Let I_0 be an idea with novelty $N(I_0, t_0) > 0$ published at t_0, and let I_1 be a reformulation with $N(I_1, t_1) = 0$ (no additional novelty) published at $t_1 > t_0$. Suppose ideas compete for adoption: each agent adopts at most one version. Let $\bar{K} = \mathbb{E}_\alpha[K(I_0 \mid S_\alpha)]$ be the average integration cost of I_0, N the total agent population, and $n_0 = |\{\alpha : I_0 \in S_\alpha(t_1)\}|$ the number of agents who have already adopted I_0 at the time I_1 appears. If the compression contribution satisfies*

$$C(I_1, I_0) > \bar{K} \cdot \frac{\ln(n_0)}{\ln(N)} \tag{8}$$

then I_1 reaches majority adoption ($N/2$ agents) before I_0 does. In a competitive regime where the version that reaches majority first captures the population, the reformulation displaces the original despite containing zero additional novelty.

Proof. Model idea propagation as a continuous-time process. Let $n_i(t)$ denote the number of agents who have adopted I_i at time t. By Proposition 1, the adoption rates follow logistic dynamics:

$$\frac{dn_0}{dt} = \lambda_0 \cdot n_0(t) \cdot (N - n_0(t)), \qquad \frac{dn_1}{dt} = \lambda_1 \cdot n_1(t) \cdot (N - n_1(t)) \tag{9}$$

where $\lambda_i \propto V(I_i)/\mathbb{E}[K(I_i \mid S_\alpha)]$ is the effective transmission rate and N is the carrying capacity.

In a fully competitive model, the dynamics are coupled (competitive Lotka–Volterra). The independent logistic approximation is conservative: in the coupled system, the faster-growing variant suppresses the slower one, strengthening our conclusion.

Since $[\![I_0]\!] = [\![I_1]\!]$, the value is identical: $V(I_0) = V(I_1)$. The transmission rates therefore satisfy:

$$\frac{\lambda_1}{\lambda_0} = \frac{\mathbb{E}[K(I_0 \mid S_\alpha)]}{\mathbb{E}[K(I_1 \mid S_\alpha)]} = \frac{\bar{K}}{\bar{K} - C(I_1, I_0)} \tag{10}$$

In this logistic model, the time to reach $N/2$ adoption from initial adopter count n is $T_{1/2}(n) = \frac{1}{\lambda N} \ln\left(\frac{N-n}{n}\right)$. For I_1 (starting from $n_1(t_1) = 1$) to reach half-adoption before I_0 (starting from n_0):

$$\frac{1}{\lambda_1 N} \ln(N-1) < \frac{1}{\lambda_0 N} \ln\left(\frac{N - n_0}{n_0}\right) \tag{11}$$

The N cancels. For $N \gg n_0 \gg 1$, we approximate $\ln(N-1) \approx \ln N$ and $\ln\left(\frac{N-n_0}{n_0}\right) \approx \ln(N/n_0)$:

$$\frac{1}{\lambda_1} \ln(N) < \frac{1}{\lambda_0} \ln(N/n_0) \tag{12}$$

Substituting $\lambda_1/\lambda_0 = \bar{K}/(\bar{K} - C)$:

$$\frac{\bar{K} - C}{\bar{K}} < \frac{\ln(N/n_0)}{\ln(N)} \tag{13}$$

Rearranging:

$$C > \bar{K}\left(1 - \frac{\ln(N/n_0)}{\ln(N)}\right) = \bar{K} \cdot \frac{\ln(n_0)}{\ln(N)} \tag{14}$$

Note that $\ln(n_0)/\ln(N) < 1$ whenever $n_0 < N$, so the required compression is strictly less than $\bar{K}$. For early-stage ideas with small n_0, the threshold is low: if $n_0 = N^{0.1}$, the condition becomes $C > 0.1\,\bar{K}$, a 10% reduction in integration cost. Compression advantages compound exponentially in the logistic regime, making the reformulation's eventual dominance robust to perturbations. □

Corollary 3 (Inevitability of Compression Credit)**.** *In a competitive adoption regime, for any fixed $n_0 > 0$ (head start of the original idea) and any compression contribution $C > 0$, there exists a population size N^* such that for all $N > N^*$, the reformulation reaches majority adoption before the original.*

Proof. By Theorem 2, the required compression threshold is $\bar{K} \cdot \ln(n_0)/\ln(N)$. For fixed n_0 and $\bar{K}$, this threshold decreases monotonically as N grows and approaches 0 as $N \to \infty$. For any $C > 0$, there exists $N^* = n_0^{\bar{K}/C}$ such that $C > \bar{K} \cdot \ln(n_0)/\ln(N)$ for all $N > N^*$. That is, in a sufficiently large civilization, even a marginal compression advantage is enough to displace the original. □

3.2 The Einstein–Poincaré Theorem

We now formalize the specific case where the reformulation also eliminates unnecessary structure.

Definition 6 (Structural Compression)**.** *Let idea I_0 contain both essential content E and auxiliary scaffolding A (theoretical commitments not required by the formalism), so that $I_0 = (E, A)$. A* structural compression *is a reformulation $I_1 = E'$ where E' is functionally equivalent to E but does not require A.*

Theorem 4 (Structural Compression Amplifies Memetic Fitness)**.** *If $I_0 = (E, A)$ and $I_1 = E'$ with $[\![E]\!] = [\![E']\!]$, then:*

$$C(I_1, I_0) \geq \mathbb{E}_\alpha[K(A \mid S_\alpha) - \delta(\alpha)] - K(E' \mid E) - O(\log k) \tag{15}$$

where $\delta(\alpha) = K(E \mid S_\alpha) - K(E \mid S_\alpha, A) \geq 0$ is the scaffolding utility for agent α (how much knowing A helps in understanding E), and $k = \max_\alpha K(A \mid S_\alpha)$ absorbs the logarithmic overhead from the chain rule. The compression contribution is the average cost of the scaffolding, reduced by its average utility and the cost of reformulation.

Proof. For any agent α, the chain rule for Kolmogorov complexity gives:

$$K(I_0 \mid S_\alpha) = K(E, A \mid S_\alpha) \tag{16}$$
$$\geq K(E \mid S_\alpha, A) + K(A \mid S_\alpha) - O(\log K(A \mid S_\alpha)) \tag{17}$$

Meanwhile, since E' is computable from E:

$$K(I_1 \mid S_\alpha) = K(E' \mid S_\alpha) \leq K(E' \mid E) + K(E \mid S_\alpha) \tag{18}$$

Taking the difference:

$$K(I_0 \mid S_\alpha) - K(I_1 \mid S_\alpha) \geq K(A \mid S_\alpha) + K(E \mid S_\alpha, A) - K(E \mid S_\alpha) - K(E' \mid E) - O(\log K(A \mid S_\alpha)) \tag{19}$$
$$= K(A \mid S_\alpha) - \delta(\alpha) - K(E' \mid E) - O(\log K(A \mid S_\alpha)) \tag{20}$$

where $\delta(\alpha) = K(E \mid S_\alpha) - K(E \mid S_\alpha, A) \geq 0$ since additional information cannot increase Kolmogorov complexity. Taking expectation over α yields the stated bound. The bound is strongest when the scaffolding A is truly unnecessary for understanding E (so $\delta \approx 0$), and weakest when A is essential for understanding E (large δ). □

Remark 2 (Application to Einstein)**.** *Poincaré's special relativity was $I_0 = (E, A)$ where E included the Lorentz transformations, time dilation, length contraction, and the relativity principle, and A was the luminiferous ether as a privileged reference frame. Einstein's special relativity was, to first approximation, $I_1 = E'$ with A removed: the same kinematic predictions without the ether. The ether contributed $K(A)$ bits of structure that the formalism did not require, and that many receiving agents did not have in S_α to begin with (the ether was increasingly suspect post-Michelson–Morley). Einstein's structural compression eliminated A, achieving $C > 0$ on this axis.*

This is not the whole story. Einstein's 1905 papers also contributed real novelty above Lorentz–Poincaré: a fully relational kinematic interpretation in which neither space nor time is privileged, the abandonment of absolute simultaneity (which Poincaré continued to entertain), and mass–energy equivalence in a companion paper. So $N(I_1, 1905) > 0$, not zero. The case is best read as one where both C and N are positive, with C large enough relative to the original $\bar{K}$ that Theorem 2 fires regardless of where the small additional N lies. The structural-compression channel is what the present theorem isolates; it is sufficient for the conclusion without claiming Einstein added nothing.

3.3 The Counterfactual Delay Function

We now address the question: should credit attribution track novelty, or does it track memetic fitness?

Definition 7 (Counterfactual Delay)**.** *The counterfactual delay $\Delta(I, t)$ of idea I published at time t is:*

$$\Delta(I, t) = \inf\{t' > t : K([\![I]\!] \mid S_C(t') \setminus \{I\}) = 0\} - t \tag{21}$$

the time it would take for the same functional content to be independently rediscovered if I had not been published.

Proposition 5. *The counterfactual delay is bounded by:*

$$\Delta(I, t) \leq \frac{2^{K(I|S_C(t))}}{\rho(t)} \tag{22}$$

where $\rho(t) = \frac{d}{dt}(\sum_\alpha |S_\alpha(t)|)$ is the aggregate rate of knowledge production across the civilization.

Proof. The community is exploring program space at aggregate rate ρ. By Levin's universal search [Levin, 1973], the expected time to find a program of length ℓ is proportional to 2^ℓ, since the algorithmic probability of a random program having a given ℓ-bit prefix is $2^{-\ell}$. A program of conditional complexity $K(I \mid S_C) = \ell$ bits will therefore be encountered within $2^\ell/\rho$ time in expectation. Ideas with low conditional complexity (small ℓ) are rediscovered exponentially faster than ideas with high conditional complexity, which is consistent with the empirical observation that ripe ideas are often rediscovered independently by multiple groups within a short window. □

Remark 3 (Ripe for Discovery)**.** *Einstein himself wrote in 1953:* There is no doubt that the special theory of relativity, if we regard its development in retrospect, was ripe for discovery in 1905 *[Einstein, 1953]. In our framework, this translates to: $K([\![I_{\text{SR}}]\!] \mid S_C(1905))$ was very small. The Lorentz transformations existed, the relativity principle was articulated, the experimental anomalies were known. The counterfactual delay Δ was short, perhaps a few years. Schmidhuber's 1991 FWP had $N(I, 1991) > 0$ but the transformer architecture at scale would likely have been developed within a few years of the attention mechanism becoming popular regardless, suggesting moderate Δ for the full* Attention Is All You Need *package.*

3.4 Compression and Novelty are Different Problems

A natural objection is that compression is a lesser intellectual activity than frontier discovery. The framework above does not show this. What it does show is that compression and novelty are distinct optimization problems with different objectives and different constraint structures.

Proposition 6 (Distinct Objectives)**.** *Let $\mathcal{F}$ denote the frontier of unsolved problems. Novelty generation is a search over program space $\{0,1\}^*$ for a program p satisfying $U(p) \in \mathcal{F}$, with acceptance determined only by the problem domain. Compression of an existing idea I_0 for a population $\mathcal{A}$ is a search for a program p' minimizing $\mathbb{E}_{\alpha\sim\mathcal{A}}[K_{U_\alpha}(p' \mid S_\alpha)]$ subject to $[\![U(p')]\!] = [\![I_0]\!]$, with acceptance determined by an average over agent-specific knowledge states and effective UTMs. These problems have different objective functions, different feasibility constraints, and in general different optimizers.*

The two problems are intractable in different ways. Novelty search is bounded below by Levin search at $\Omega(2^{K(I|S_C)})$ when no closer prior is available. Compression search has an immediate feasible solution (republish I_0 verbatim) and must therefore be evaluated by how much improvement over that baseline is achievable; the optimal compression is also uncomputable in the limit because K is uncomputable. Neither problem is uniformly easier than the other.

What the proposition does establish is that a researcher optimizing one objective is not automatically optimizing the other. The empirical pattern that the same individual rarely contributes both the foundational novelty and the dominant reformulation of the same idea is consistent with this. Einstein's two-postulate derivation was not a simplification in the trivial sense; it was a solution to a different problem from the one Lorentz and Poincaré were solving. Treating compression as a lesser activity collapses this distinction.

4 Machine Civilizations

4.1 The Attribution Problem Persists

One might expect that machine civilizations, with perfect information and no social biases, would solve the credit attribution problem. We show this is not the case.

Theorem 7 (Attribution Bias is Bandwidth-Intrinsic)**.** *In any civilization of N agents optimizing for aggregate knowledge, a contribution has two channels of value: (i)* frontier value*, the novelty $N(I)$ that pushes the boundary of what is known (realized once), and (ii)* diffusion value*, the compression $C(I)$ that reduces the cost of propagating the result to N agents (realized N times). The equilibrium credit weight assigned to the compression contribution is:*

$$w_{\text{compress}} = \frac{N \cdot C(I)}{N \cdot C(I) + N(I)} \tag{23}$$

As $N \to \infty$, $w_{\text{compress}} \to 1$ for any $C > 0$.

Proof. Consider a machine civilization optimizing for aggregate knowledge growth $\frac{d}{dt}\sum_\alpha |S_\alpha(t)|$. An idea I contributes to this sum in two ways. First, I expands the frontier of computable knowledge by $N(I)$: this is a one-time gain independent of N. Second, each agent that integrates I gains value $V(I)$, and the cost of integration is $K(I \mid S_\alpha)$. A reformulation I_1 with $C(I_1, I_0) > 0$ reduces this per-agent cost, so the aggregate savings across N agents are:

$$U_{\text{compress}} = N \cdot C(I_1, I_0) \tag{24}$$

while the frontier value remains:

$$U_{\text{novel}} = N(I_0) \tag{25}$$

A rational credit allocation weights each type of contribution by its total utility:

$$w_{\text{compress}} = \frac{U_{\text{compress}}}{U_{\text{compress}} + U_{\text{novel}}} = \frac{N \cdot C}{N \cdot C + N(I_0)} \tag{26}$$

which is bounded in $[0, 1]$ and increases monotonically in N. □

Remark 4 (Multiplicative novelty effects)**.** *The proof treats $U_{\text{novel}} = N(I)$ as independent of N. One might object that novelty has downstream multiplier effects: a frontier-expanding idea enables further discoveries by all N agents, so its effective value should scale as $N \cdot g(N)$ for some increasing function g. If $g(N) = N^\beta$ for $\beta \in [0, 1]$, the corrected weight becomes:*

$$w_{\text{compress}} = \frac{N \cdot C}{N \cdot C + N^\beta \cdot N(I)} \tag{27}$$

For $\beta < 1$ (sublinear multiplier effects), $w_{\text{compress}} \to 1$ as $N \to \infty$ still holds. The critical boundary is $\beta = 1$: if novelty value scales linearly with N, then $w_{\text{compress}} = C/(C + N(I))$, a constant independent of N. The empirical observation that most ideas are not built upon by most agents suggests $\beta < 1$ in practice, preserving the main result.

Corollary 8 (Machine Civilizations Amplify Compression Credit)**.** *A machine civilization with large N (e.g., a galaxy-spanning civilization with light-speed communication limits) will allocate nearly all credit to compression contributions, not novelty contributions. For $N = 10^6$ and $C/N(I) = 0.01$ (a reformulation whose compression gain is 1% of the original novelty), the compression weight is already $w_{\text{compress}} \approx 0.9999$.*

4.2 Intractability of Novelty-Weighted Correction

A defender of novelty-weighted credit might accept the above and respond: *We should maintain a provenance graph and reallocate credit to originators, correcting for the compression bias.* We prove this correction is computationally intractable and that any civilization expending resources on it is provably slower.

Theorem 9 (Intractability of Credit Correction)**.** *Let $\mathcal{C}$ denote a credit-correction mechanism that, for each adopted idea I_1, computes the true novelty contribution of all prior ideas $\{I_0^{(1)}, \ldots, I_0^{(m)}\}$ in the provenance chain. Computing $N(I_0^{(j)}, t_0^{(j)})$ requires evaluating the counterfactual $K(\Omega_D \mid S_C(t_0^{(j)}) \setminus \{I_0^{(j)}\})$, which is at least as hard as re-running the original discovery process. Specifically:*

(i) Evaluating the counterfactual delay $\Delta(I, t)$ for a single idea requires simulating the civilization's search process without I, at cost $\Omega(2^{K(I|S_C(t))})$.

(ii) For a knowledge base of M ideas, maintaining exact novelty-weighted credit requires $O(M)$ such evaluations per new adoption event.

(iii) A civilization that allocates fraction f_{correct} of its compute budget to credit correction reduces its rate of intellectual progress by factor $(1 - f_{\text{correct}})$ while the correction provides zero bits of new knowledge or new compression. The steady-state growth rate satisfies:

$$G_{\text{corrected}} = (1 - f_{\text{correct}}) \cdot G_{\text{uncorrected}} \tag{28}$$

Therefore any $f_{\text{correct}} > 0$ is Pareto-dominated by $f_{\text{correct}} = 0$.

Proof. For (i): Computing $N(I_0, t_0)$ requires evaluating $K(\Omega_D \mid S_C(t_0))$ and $K(\Omega_D \mid S_C(t_0) \cup \{I_0\})$, then taking the difference. The first term requires determining what the civilization could compute without I_0, which amounts to simulating the civilization's search trajectory on the counterfactual branch. By the halting problem, determining whether a given program contributes to solving problems in Ω_D is undecidable in general; the best computable approximation is Levin search, requiring $\Omega(2^{K(I|S_C)})$ steps.

For (ii): Each new idea I_1 potentially changes the novelty attribution of every prior idea in its provenance chain (because the counterfactual of removing $I_0^{(j)}$ now includes a world where I_1 might have been discovered via a different path). This creates cascading recomputation: adopting one idea can invalidate the credit assignments of all preceding ideas.

For (iii): Credit correction produces a mapping from ideas to real-valued novelty scores. This mapping has positive Kolmogorov complexity (it must be stored and maintained) but contributes nothing to frontier expansion and nothing to compression of existing knowledge. Every compute cycle spent on correction is a cycle not spent on search or compression. Since both numerator (useful work) and denominator (total compute) change by the same factor, and the correction adds zero to the numerator, the growth rate drops by exactly $(1 - f_{\text{correct}})$. □

Remark 5 (The irony of advocacy)**.** *A researcher who spends time publicly advocating for retroactive credit reallocation is, by this theorem, spending resources on credit correction rather than either novelty generation or compression. The theorem predicts that this is the lowest-value activity available to such a researcher. The time would be better spent either (a) generating new*

ideas or (b) compressing existing ideas for broader audiences, that is, doing what the theory says matters. The researcher's own most impactful contributions will be those where compression was done well, not those where priority was established first.

4.3 The Optimal Search Strategy

Theorem 10 (Optimal Intellectual Investment)**.** *Consider a civilization of N agents optimizing for rate of progress, with two scaling assumptions: (i) the rate of novel idea generation by k agents scales as $\rho(k) = c\sqrt{k}$ for some constant $c > 0$, reflecting diminishing returns on parallel search in a fixed problem space; (ii) aggregate communication bandwidth scales as $W(N) = aN$ for some constant $a > 0$, reflecting constant per-agent outgoing capacity. Let $\bar{K}$ denote the average per-agent cost to integrate one idea. The optimal allocation satisfies:*

$$c\sqrt{f_{\text{novel}} \cdot N} = (1 - f_{\text{novel}}) \cdot \frac{a}{\bar{K}} \tag{29}$$

which for large N gives $f_{\text{novel}} \sim \frac{a^2}{c^2 \bar{K}^2 N} \sim \frac{1}{N}$, with $f_{\text{compress}} = 1 - f_{\text{novel}}$.

Proof. Allocate fraction f_{novel} of agents to novelty search and $f_{\text{compress}} = 1 - f_{\text{novel}}$ to compression and communication. By the diminishing-returns assumption, $f_{\text{novel}} \cdot N$ agents searching in parallel generate novel ideas at rate $r_n = c\sqrt{f_{\text{novel}} \cdot N}$. The remaining agents provide compression bandwidth $(1 - f_{\text{novel}}) \cdot aN$; each idea costs $N \cdot \bar{K}$ bandwidth to propagate to the full population, so ideas are processed at rate $r_c = (1 - f_{\text{novel}}) \cdot a/\bar{K}$. Aggregate knowledge growth is bottlenecked by the slower stage:

$$G = \min(r_n, r_c) \cdot N \cdot \bar{V} \tag{30}$$

G is maximized when the pipeline is balanced: $r_n = r_c$. Setting them equal:

$$c\sqrt{f_{\text{novel}} \cdot N} = (1 - f_{\text{novel}}) \cdot \frac{a}{\bar{K}} \tag{31}$$

For large N, $f_{\text{novel}} \ll 1$, so $(1 - f_{\text{novel}}) \approx 1$. Squaring both sides:

$$c^2 f_{\text{novel}} N \approx \frac{a^2}{\bar{K}^2} \tag{32}$$

Solving:

$$f_{\text{novel}} \approx \frac{a^2}{c^2 \bar{K}^2 N} \tag{33}$$

which scales as $1/N$. The diminishing-returns exponent in ρ propagates squared into the allocation: because the search rate scales as $\sqrt{k}$, halving the searcher count reduces the rate by only $\sqrt{2}$, so the optimizer can shrink the search pool aggressively. □

Remark 6 (Scaling assumptions are load-bearing)**.** *The specific $1/N$ exponent follows directly from two choices: square-root diminishing returns on parallel search ($\rho(k) \propto k^{1/2}$) and linear bandwidth scaling ($W(N) \propto N$). Both are convenient parameterizations, not measured laws. If $\rho(k) \propto k^{\gamma}$, the optimal allocation scales as $f_{\text{novel}} \sim N^{-1/(2(1-\gamma))}$, which for $\gamma \to 1$ (no diminishing returns) vanishes only as a constant fraction and for $\gamma \to 0$ (sharp diminishing returns) vanishes faster than $1/N$. If bandwidth grows sublinearly in N (a realistic case at galactic scale due to light-speed constraints), the optimal novel fraction shrinks further.*

What is robust across choices is the qualitative claim: under any combination of diminishing returns on parallel search and finite aggregate bandwidth, compression and communication eventually dominate novelty in the optimal resource allocation as N grows. The specific exponent should be read as the behaviour under one tractable parameterization, not as a universal law. The empirical observation that mature scientific fields spend most of their effort on synthesis, replication, extension, and communication rather than on fundamental discovery is consistent with any plausible parameterization of this kind.

5 Case Studies

5.1 Lorentz–Poincaré–Einstein

In our framework:

Lorentz (1892–1904): High novelty N. Derived the transformation equations from electrodynamic first principles. But high conditional complexity $K(I \mid S_\alpha)$ for non-specialists: required deep familiarity with ether physics.

Poincaré (1898–1905): Moderate additional novelty (group structure, principle of relativity, clock synchronization). But retained the ether as scaffolding A, keeping $K(I_0 \mid S_\alpha)$ high for agents without ether-theoretic training, and continued to entertain absolute simultaneity.

Einstein (1905): Small but nonzero additional novelty (fully relational kinematics with neither space nor time privileged, formal abandonment of absolute simultaneity, mass–energy equivalence in a companion paper). And large compression contribution C through structural compression: eliminated the ether (removed scaffolding A), derived everything from two postulates, used thought experiments that minimized $K(I \mid S_\alpha)$ for broad agent populations. The case is best modelled as $N > 0$ small and C large enough that Theorem 2 fires regardless.

Einstein's memetic fitness F dominated because he minimized integration cost across the widest possible agent distribution. His thought experiments (trains, elevators, light beams) were compression devices: they reduced $K(I \mid S_\alpha)$ for agents α with no background in electrodynamics. The 1919 Eddington eclipse expedition functioned as a *bandwidth amplifier*, although a subtle one: the eclipse confirmed a general-relativity prediction (gravitational light bending), not a special-relativity prediction. The general public made no such distinction, so Eddington's confirmation amplified Einstein's overall public profile and thereby retroactively boosted SR adoption as well. The amplification is real; the channel is GR, not SR. A more careful operationalization would treat Eddington as raising $V(I, \alpha)$ for the public by associating Einstein's name with a dramatic experimental result, with the integration-cost reduction following indirectly.

5.2 Schmidhuber–Vaswani

Schmidhuber (1991–1993): High novelty $N > 0$. First end-to-end differentiable fast weight programmer using outer products (keys/values) with gradient descent training. Published with equations, experiments, and extensions to recurrence and self-reference. But high $K(I \mid S_\alpha)$: the fast weight framing was unfamiliar to most practitioners; the tech report format limited distribution; the 1991 compute environment meant experiments were necessarily toy-scale, providing no empirical evidence of the mechanism's power at useful problem sizes.

Vaswani et al. (2017): Some additional novelty (softmax attention, multi-head projections, positional encodings, the specific encoder-decoder Transformer architecture for sequence-to-sequence translation) and large compression contribution on multiple axes. The novelty is not zero: Schlag et al. [2021] proved that *linear* Transformers (attention without softmax) are mathematically equivalent to FWP, but the 2017 paper used softmax attention, which empirically outperforms the linear variant and is what subsequent practice settled on. So the equivalence is between FWP and one variant of the Transformer family, not between FWP and the Transformer as deployed.

The compression contribution is what does most of the work. The name *Transformer*, the clean implementation, and Google's institutional bandwidth all reduced $K(I_1 \mid S_\alpha)$ for the average 2017 ML practitioner. By 2017, GPU compute had become several orders of magnitude cheaper than in 1991, enabling large-scale experiments that demonstrated the mechanism's practical value. The dramatic machine translation results changed $V(I_1, \alpha)$: agents could now see that the mechanism worked at scale, collapsing the integration cost from *understand and trust an untested formalism* to *replicate a demonstrated result.* Cheap compute acted as an

exogenous reduction in $K(I \mid S_\alpha)$ across the entire agent population, a form of environmental compression that the 2017 paper could exploit and the 1991 paper could not.

The credit gap between Vaswani 2017 and Schmidhuber 1991 is therefore mostly explained by compression and environmental factors, with a smaller real-novelty component from softmax attention and the specific architectural assembly. This is consistent with Theorem 2: C does not have to account for 100% of the gap for the reformulation to dominate, only enough to clear the threshold $\bar{K} \cdot \ln(n_0)/\ln(N)$, which for the 2017 ML community was very low.

5.3 LSTM: The Confirming Case

The LSTM case is instructive precisely because it is the exception that confirms the rule. Hochreiter and Schmidhuber published Long Short-Term Memory in 1997 [Hochreiter and Schmidhuber, 1997]. Unlike the 1991 FWP, LSTM exhibited high compression contribution from the start:

Naming: *Long Short-Term Memory* is a memorable, self-explanatory name. Compare *Fast Weight Programmer*, which requires understanding what fast weights are before the name conveys meaning.

Framing: The vanishing gradient problem was well-known by 1997 and widely felt as a pain point. LSTM was framed as a solution to an existing, recognized problem, not as a novel paradigm. This reduced $K(I \mid S_\alpha)$ dramatically: agents did not need to understand *why* they needed the mechanism, only *how* it worked.

Empirical accessibility: While 1997 compute was still limited, LSTM operated on tasks (sequence prediction, simple language modeling) that were within the compute budget of the average researcher. Replication was feasible, unlike the FWP's more abstract demonstrations.

The result: LSTM credit allocation tracks the theory's predictions. Hochreiter and Schmidhuber receive substantial credit for LSTM. The credit gap between LSTM and its precursors is small compared to the credit gap between Transformers and the FWP. The difference is not that the community was fairer in one case and biased in the other. The difference is that LSTM was compressed well and the FWP was not.

By Proposition 6, this is not a lesser achievement, only a different one. The compression optimization that produced LSTM's clean problem-solution framing is a distinct problem from the novelty search that produced the FWP's core mechanism, and the empirical pattern across Schmidhuber's publication record is consistent with the model: impact tracks compression contribution closely, and most of the variance across his outputs is explained by how well each result was packaged for the receiving audience of its era.

5.4 Quantitative Comparison

Let us estimate the key quantities for all three cases. The following are illustrative order-of-magnitude estimates, not empirical measurements; they serve to demonstrate that the framework produces qualitatively correct predictions about relative credit allocation.

Quantity	Einstein/Poincaré	Vaswani/Schmidhuber	LSTM
Novelty ratio $N(I_1)/N(I_0)$	≈ 0.05	≈ 0.15	≈ 0.9
Compression ratio $C/\bar{K}$	≈ 0.6	≈ 0.7	≈ 0.5
Counterfactual delay $\Delta(I_0)$	2–5 years	5–15 years	3–8 years
Counterfactual delay $\Delta(I_1)$	1–2 years	1–3 years	N/A
Public credit ratio $I_1 : I_0$	$\approx 95 : 5$	$\approx 98 : 2$	$\approx 40 : 60$

The LSTM column is revealing. The novelty ratio is high (LSTM is largely attributable to its named authors), the compression ratio is moderate (good but not extraordinary), and the

public credit allocation is roughly proportional. When compression and novelty are contributed by the same authors, the credit allocation is approximately fair. The injustice arises only when different agents contribute novelty and compression, and the theory explains why: it is not an injustice but an accurate accounting of total value, where compression value scales with population and novelty value does not.

The counterfactual delay for the 1991 FWP is arguably longer than for special relativity: in 1905, multiple physicists were converging on the same results. In 1991, the FWP concept was far ahead of its time, and no one else was working on gradient-descent-trained outer-product fast weight memories. This suggests that by a novelty-weighted counterfactual metric, Schmidhuber's credit deficit is actually larger than Poincaré's.

6 A Conditional Extension: Credit in Self-Improving Systems

We now consider a conditional extension of the framework. Suppose a self-improving system (a Gödel Machine [Schmidhuber, 2006] that can rewrite its own code including its credit-assignment mechanisms) adopts compression progress as its sole optimization target at the meta-level. What does the resulting attribution behaviour look like?

We emphasize at the outset that compression progress was originally proposed by Schmidhuber [2009] as an intrinsic-motivation signal for individual artificial agents, not as a meta-principle for organizing the social process of scientific credit. The construction in this section is a thought experiment: *if* such an objective were extended to govern attribution in an artificial scientific community, *then* the following holds. Whether this extension is desirable, or whether human research communities should adopt anything like it, is a separate question we do not address.

Definition 8 (Compression Progress)**.** *Following Schmidhuber [2009], define the compression progress at time t as:*

$$CP(t) = K(H_{<t} \mid S(t-1)) - K(H_{<t} \mid S(t)) \tag{34}$$

where $H_{<t}$ is the history of observations up to t and $S(t)$ is the system's model at time t.

A self-improving system maximizing CP would: (1) maintain a provenance graph for debugging and reliability, not for credit; (2) allocate search effort to the type of contribution (novelty vs. compression) that maximizes expected CP per unit compute; (3) by Theorem 10, allocate $\sim 1 - 1/N$ of resources to compression as the system scales.

Theorem 11 (Attribution Metadata is Optimally Discarded)**.** *Let $S(t)$ be the knowledge state of a self-improving system maximizing compression progress, and let $A(t)$ be the attribution metadata at time t (a mapping from each idea $I \in S(t)$ to its provenance: originator, timestamp, novelty score). Then:*

(i) $K(A(t)) > 0$ *for any non-trivial knowledge base.*

(ii) $CP_{\text{attributable}} = K(H_{<t} \mid S(t) \setminus A(t)) - K(H_{<t} \mid S(t)) = 0$*, that is, attribution metadata contributes zero compression progress.*

(iii) A system maximizing CP subject to a storage constraint will discard $A(t)$ before discarding any predictively useful component of $S(t)$.

Proof. For (i): The attribution mapping $A(t)$ assigns to each idea a tuple (originator ID, timestamp, novelty score). For M ideas, this requires $\Omega(M \log M)$ bits (at minimum, a permutation over idea-originator pairs). Thus $K(A(t)) > 0$ for $M > 1$.

For (ii): Attribution metadata records *who* produced an idea, not *what* the idea predicts. The compression progress $CP(t)$ depends only on how well $S(t)$ compresses the observation

history $H_{<t}$. Since $H_{<t}$ contains observations about the world, not about the social process of idea generation, $K(H_{<t} \mid S(t)) = K(H_{<t} \mid S(t) \setminus A(t))$: removing $A(t)$ does not change the system's ability to compress observations. (If $H_{<t}$ *does* contain observations about the social process, then attribution metadata becomes a predictive model of social dynamics and is no longer metadata but science, which should be retained on its merits.)

For (iii): Under a storage constraint $|S(t)| \leq B_S$, the system must choose which components to retain. Any component C with $CP_C > 0$ (contributes to compression progress) dominates any component with $CP_C = 0$ (does not contribute). Since $CP_{A(t)} = 0$ and $K(A(t)) > 0$, discarding $A(t)$ frees $K(A(t))$ bits for components with positive CP, strictly increasing the system's compression rate. □

The theorem is conditional. *If* a system maximizes compression progress as its sole objective, *then* it provably discards attribution metadata. The theorem does not say that compression progress *should* be the objective of any human or artificial research community. It says only that a community organized around this particular objective would not maintain priority records.

Under this conditional, the distinction between *who thought of it first* and *who wrote it most efficiently* becomes invisible in the maximally compressed representation, because that representation retains only the bits that contribute to compressing future observations. Attribution is overhead with respect to that specific objective. A community optimizing this specific objective would shed it.

This is one possible meta-architecture among many. A community that valued provenance independently could simply add attribution to its objective function, paying the storage cost knowingly. The result of this section is therefore a tradeoff statement, not a recommendation.

7 Limitations

The framework is stylized in several ways that bear on how its conclusions should be read.

Kolmogorov complexity is uncomputable. The integration cost $K(I \mid S_\alpha)$ and the novelty contribution $N(I, t)$ are mathematically well-defined but not measurable in practice. The case studies use the formalism heuristically: we identify situations where the qualitative ordering of K values across reformulations is plausible (Einstein's reformulation dropped scaffolding bits relative to Poincaré's; the 2017 Transformer paper benefited from cheaper compute and clearer naming relative to the 1991 FWP) without claiming numerical estimates. Any quantitative use of the framework would require committing to a particular approximation scheme (e.g., compressor-based estimates, language-model probabilities) and would inherit that scheme's biases.

Operationalization of memetic dynamics is heuristic. Real-world adoption depends on factors the model does not explicitly distinguish: trust in the author, peer effects, the host institution's status, the language of publication, the political and economic context, and the timing relative to enabling technologies. The model treats some of these as bandwidth amplification but does not separate them from compression proper. A more complete theory would decompose the value $V(I, \alpha)$ and the effective UTM U_α into these components and study their relative weights empirically.

Scaling assumptions in Section 3.4 and Theorem 10 are toy choices. The square-root diminishing-returns law for parallel search and the linear bandwidth scaling are convenient parameterizations, not measured laws. Different choices give different exponents in the optimal allocation. The qualitative claim that compression eventually dominates novelty in resource allocation as N grows is robust to a range of choices; the specific $1/N$ exponent is not.

The framework is descriptive, not normative. The theorems predict the allocation that emerges under bandwidth constraints in our model. They do not establish that this allocation is fair or correct under any particular normative criterion. Two reasonable normative

positions are compatible with the same descriptive results: (i) credit ought to track adoption-weighted impact, in which case the system's allocation is correct; (ii) credit ought to track counterfactual displacement, in which case the system's allocation is biased and the bias is a property worth acknowledging even if correcting it is intractable. The paper does not adjudicate between these positions.

The self-improving extension is a thought experiment. Section 6 treats the case where compression progress is the sole optimization target of an artificial scientific community. This is a conditional construction, not a claim about what human research communities should do or what Schmidhuber [2009] proposed.

The logistic adoption model is a simplification. Theorem 2 uses independent logistic dynamics; the coupled Lotka–Volterra version strengthens the conclusion but the simplifying assumption that $V(I_0) = V(I_1)$ for two reformulations is itself an idealization, since reformulations can change perceived value as well as integration cost. The Schmidhuber–Vaswani case study explicitly uses the latter mechanism (compute availability changing V) which our theorem does not formally handle.

We list these limitations because the rhetorical force of a formal framework can outrun its actual reach. The framework is most safely read as a model of one mechanism (compression-weighted adoption) operating alongside others, not as a complete theory of scientific credit.

8 Conclusion

We have shown the following, subject to the model assumptions made along the way:

1. Adoption-weighted credit allocation in science tracks memetic fitness, which is dominated by compression contribution rather than novelty contribution in the model (Theorem 2).

2. Structural compression—eliminating unnecessary theoretical scaffolding—amplifies memetic fitness through a chain-rule decomposition, consistent with the empirical fact that Einstein's ether-free reformulation captured credit from the mathematically close Lorentz–Poincaré apparatus (Theorem 4).

3. Compression and novelty are distinct optimization problems with different objectives and constraint structures, and neither is uniformly easier than the other (Proposition 6).

4. Under the bandwidth-budget accounting of Theorem 7, the equilibrium credit weight assigned to compression grows with population size; machine civilizations of large N would exhibit this allocation more strongly, not less.

5. Attempting to maintain exact novelty-weighted credit at the meta-level is computationally intractable and reduces the rate of intellectual progress under the cost model of Theorem 9.

6. Under a tractable parameterization, a civilization optimizing for aggregate knowledge growth allocates a fraction of effort to novelty that decreases with N (Theorem 10). The specific $1/N$ exponent is parameterization-dependent; the qualitative dominance of compression at scale is robust across reasonable choices.

7. Conditional on a community adopting compression progress as its sole meta-objective, attribution metadata is provably discarded as zero-value overhead (Theorem 11). This is a conditional result, not a recommendation.

The descriptive implication for researchers: the community's allocation of credit to Einstein over Poincaré, or to the Transformer paper over the 1991 FWP, is the allocation that bandwidth-constrained adoption dynamics predict. Whether this allocation is *fair* in any normative sense

depends on a premise the framework does not argue for: that credit ought to track adoption-weighted impact rather than counterfactual displacement. The descriptive claim is robust within the model; the normative one is a choice.

The actionable corollary is independent of the normative question. Inventing without compressing leaves most of the diffusion value on the table, and Theorem 9 shows that fighting the resulting allocation through retroactive correction is costly under the model. Compression for a heterogeneous audience is a different problem from frontier discovery, not a lesser one (Proposition 6), and a researcher who optimizes only one of the two is leaving the other gain on the table.

The LSTM case demonstrates the model's behaviour when the same researcher contributes both novelty and compression: the allocation looks close to fair on a novelty metric and close to fair on a fitness metric simultaneously, because the two contributions are not separated across authors. When they are separated, the metrics diverge, and the fitness metric is what the system tracks.

References

S. I. Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions*, C-21:1197–1206, 1972.

Albert Einstein. Letter to C. Seelig, 1953. Quoted in A. Pais, *Subtle is the Lord*, Oxford University Press, 1982.

Donald O. Hebb. *The Organization of Behavior*. Wiley, 1949.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. National Academy of Sciences*, 79(8):2554–2558, 1982.

A. G. Ivakhnenko and V. G. Lapa. *Cybernetic Predicting Devices*. CCM Information Corporation, 1965.

Jerzy Konorski. *Conditioned Reflexes and Neuron Organization*. Cambridge University Press, 1948.

Leonid A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266, 1973.

Seppo Linnainmaa. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's thesis, University of Helsinki, 1970.

Hendrik A. Lorentz. Electromagnetic phenomena in a system moving with any velocity less than that of light. *Proc. Royal Netherlands Academy of Arts and Sciences*, 6:809–831, 1904.

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63, 1968.

Henri Poincaré. Sur la dynamique de l'électron. *Comptes Rendus*, 140:1504–1508, 1905.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear Transformers are secretly fast weight programmers. In *Proc. ICML*, 2021.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to recurrent nets. Technical Report FKI-147-91, TU Munich, 1991. 26 March 1991.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

Jürgen Schmidhuber. Discovering neural nets with low Kolmogorov complexity and high generalization capability. *Neural Networks*, 10(5):857–873, 1997.

Jürgen Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial General Intelligence*, pages 199–226. Springer, 2006. Preprint 2003.

Jürgen Schmidhuber. Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of SICE*, 48(1): 21–32, 2009.

Ray J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7(1):1–22, 1964.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NIPS*, 2017.