Manufacturing Credulity

Credential Laundering, Hagiographic Genre, and Infrastructural Epistemology in the AI Public Sphere

Flyxion
Independent Researcher

April 29, 2026

Abstract. This essay proposes and theorizes the concept of *manufacturing credulity* as distinct from the better-documented phenomenon of manufacturing credibility. Where credibility manufacture concerns the construction of an appearance of trustworthiness in a source, credulity manufacture operates one level deeper: it cultivates in the audience a prior disposition to accept authoritative-sounding synthesis without demanding source constraints or internal consistency. Three instantiations of this mechanism are examined. First, the compiled lecture reel of Mo Gawdat, former Chief Business Officer of Google X, is analyzed as a recomposable fragment structure—a rhetorical form optimized for virality precisely through the decoupling of individually shareable claims from the constraints of coherent argument. Second, the hagiographic biography of Demis Hassabis is examined as a genre-level operation that launders expertise in simplified formal domains into civilizational authority, while systematically reframing structural complicity as personality texture. Third, and prefatorily, a Google AI-generated summary of the phrase "manufacturing credulity" itself—synthesized from a single Chinese-language graduate seminar abstract and presented as settled conceptual vocabulary—is treated as a real-time demonstration of credulity manufacture at infrastructural scale. These three cases are argued to be not individual failures but genre-level features of the current AI attention economy, unified by a shared deep structure: the simulation of constraint closure without the operative work of domain-specific constraint satisfaction. The essay draws on platform critique, Goodhart dynamics, proxy integrity collapse, and visibility field theory to develop this diagnosis, before arguing that genuine epistemic authority requires what these performances systematically suppress: accountability to the constraints that constitute expertise in the first place.

Prefatory Demonstration

Before the argument begins, a demonstration. In the course of researching whether the phrase "manufacturing credulity" had prior theoretical currency, a Google AI Overview was returned. It read, with full academic composure, as a synthesized definition: the phrase referred, it explained, to the deliberate and systemic process of making people willing to believe unlikely or false information, often without sufficient evidence, a concept analyzed in media studies, political science, and psychology, exploring how misinformation and propaganda cultivate a climate of uncritical acceptance. The definition was fluent, professionally hedged, and entirely plausible.

The source, displayed below the summary after some navigation, was a single result: an abstract from a graduate seminar at Peking University's School of Government, written in Chinese, concerning the convergence of China's media ecosystem and nationalism in shaping news discernment. Google's search noted, with characteristic understatement, that it had omitted results similar to the one already displayed. There was only one.

The AI had synthesized a definition of a phrase from a single non-English source and presented it in the register of established conceptual vocabulary. It performed constraint closure—the appearance of synthesis from a stable field of prior usage—where no such field existed. The failure mode is precise: a zero-shot generalization that satisfies the stylistic constraints of academic summary while violating the evidential constraints that would make such a summary legitimate. The definition it produced is not wrong, exactly, but it is not derived: it is genre-confabulation, the application of academic summary conventions to a vacuum. And it did so in response to a query about the manufacture of credulity, demonstrating the phenomenon it was asked to define.

This is the essay's entry point. What the Google summary does in miniature—simulate authoritative synthesis by suppressing source constraints, producing a readiness to believe in the absence of actual grounds—is what the AI public sphere does systematically and at scale. The following analysis examines how this operation works across three registers: the individual public intellectual, the biographical genre, and the search infrastructure itself.

Manufacturing Credulity versus Manufacturing Credibility

The phrase "manufacturing consent," introduced by Noam Chomsky and Edward Herman, describes a process by which mass media selects and frames true information to produce compliance with existing power structures. The mechanism operates on content: what is covered, how it is contextualized, whose voices are amplified. Manufacturing credibility, a phrase with existing circulation in media studies, describes a related but distinct operation at the level of source: the construction of an appearance of trustworthiness through credential display, institutional affiliation, and rhetorical authority signals.

Manufacturing credulity is neither of these, though it draws on both. It operates not on content, nor primarily on the source's apparent trustworthiness, but on the epistemic disposition of the audience prior to any specific claim. Its target is the audience's prior threshold for evidential demand—the degree of constraint satisfaction a claim must demonstrate before belief is extended—and its operation is the systematic lowering of that threshold through the accumulation of authority signals that substitute for the evidential work they appear to have already done. Where credibility manufacture asks whether the speaker seems trustworthy, credulity manufacture asks whether the audience has been conditioned to defer without noticing that it is deferring.

The distinction matters because it identifies a different kind of vulnerability. Credibility manufacture can in principle be countered by exposing the sources of authority claims—by showing that the credential is narrow, the affiliation conflicted, the track record poor. Credulity manufacture is harder to counter because it operates upstream of specific claims. An audience in a state of manufactured credulity does not evaluate individual assertions against evidence: it processes the overall texture of the performance—its fluency, its emotional arc, its apparent command of a complex domain—and extends a general warrant of belief that individual claims then draw against without individual justification.

This is why the three cases analyzed here are illuminating in combination. Each operates through a different mechanism to produce the same epistemic outcome: an audience that has extended a general warrant to a speaker, a genre, or an infrastructure, and therefore processes specific claims as confirmation rather than as propositions requiring evaluation.

Working Definitions

Three concepts recur throughout the analysis and are defined here for precision. *Manufacturing credulity* is the systematic lowering of an audience's evidential threshold through accumulated authority signals, such that subsequent claims are processed under a general warrant of belief rather than evaluated against domain-specific constraints. It is distinguished from manufacturing credibility, which targets the source, and manufacturing consent, which targets the content; credulity manufacture targets the epistemic posture of reception itself.

A *recomposable fragment* is a claim optimized for independent circulation, such that its truth conditions are evaluated locally while its contradictions with other claims in the same corpus are distributed globally and therefore remain invisible at the point of encounter. The fragment achieves shareability by decoupling from the argumentative constraints that would make its inconsistencies apparent.

Simulated constraint closure is the production of outputs that satisfy the surface regularities of a domain—its genre conventions, rhetorical markers, and stylistic norms—without satisfying the underlying admissibility conditions that constitute genuine knowledge in that domain. It is the epistemic analogue of a building facade: structurally convincing from the front, load-bearing nowhere.

Formal Analysis: Credulity as Threshold Displacement

The mechanism of credulity manufacture can be stated precisely. Let $E(c)$ denote the actual evidential support available for a claim c, and let θ be an agent's acceptance threshold, so that the agent accepts c if and only if $E(c) \geq \theta$.

Definition 2.1 (Credulity Manufacture)**.** *Credulity manufacture is a transformation $\mathcal{M}$ that produces an inflated perceived evidential score*

$$\hat{E}(c) = E(c) + \Sigma(c),$$

where $\Sigma(c) \geq 0$ represents authority signal strength—credentials, narrative fluency, institutional affiliation—such that $\hat{E}(c) \geq \theta$ while $E(c) < \theta$.

Note that $\Sigma(c)$ does not increase actual evidential support; it substitutes for it at the point of evaluation. The agent accepts c because $\hat{E}(c) \geq \theta$, but the

underlying constraint satisfaction $E(c)$ remains below threshold.

Proposition 2.2. *As authority signal strength $\Sigma(c)$ increases relative to θ, acceptance of c becomes independent of its actual evidential grounding $E(c)$.*

Proof. For any fixed θ and $E(c) < \theta$, there exists $\Sigma^* = \theta - E(c)$ such that $\hat{E}(c) = \theta$ and the agent accepts c. For $\Sigma(c) > \Sigma^*$, acceptance holds regardless of further changes to $E(c)$, provided $E(c) + \Sigma(c) \geq \theta$. In the limit $\Sigma(c) \gg \theta$, the condition $E(c) \geq 0$ suffices for acceptance, decoupling belief entirely from constraint satisfaction. □

This formalizes the distinction between manufacturing credibility—which increases $\Sigma(c)$ for a specific claim—and manufacturing credulity, which operates at the level of the general threshold θ, lowering it across a broad class of claims from a given source so that $\Sigma(c)$ need not be large for each individual claim.

The Recomposable Fragment: Mo Gawdat and the Compiled Lecture Reel

A transcript of Mo Gawdat's AI lecture reel, produced under the title "We're Entering the Most Dangerous Phase of AI Yet," reveals its production method in its own metadata. Approximately forty instances of the tag `[music]` punctuate the text, each marking a cut between fragments recorded at different times and assembled into the appearance of continuous argument. This is not a lecture. It is a highlight reel, and recognizing it as such dissolves any obligation to evaluate it as a coherent thesis.

Gawdat's biography is itself a credential-laundering operation at the origin level. He coded since age seven, joined IBM, spent eight years at Microsoft, twelve years at Google, and served as Chief Business Officer of Google X. This trajectory is presented as conferring authority over questions of AI consciousness, civilizational risk, economic transformation, relationship psychology, and ethical philosophy. The move from competence in one domain to authority in all of them is never argued for; it is performed through biographical accumulation, and the audience's credulity does the rest.

The reel's rhetorical structure instantiates what the previous section defined as the recomposable fragment architecture. Each clip makes a claim that is independently shareable and emotionally resonant: "AI already controls

our minds"; "the capitalist lie is over"; "your purpose was never to work"; "falling in love is a very complex math problem"; "I'm almost betting my life that we will see AGI in 2026." These fragments circulate independently on social media, each finding its appropriate audience: the alarmed, the spiritually restless, the economically anxious, the romantically disappointed. The contradictions between them are invisible at the fragment level and only become apparent when the whole is read consecutively, which the format is engineered to prevent. Each fragment's truth conditions are evaluated locally by whoever encounters it; the global inconsistency is nobody's problem because nobody occupies the global position.

The contradictions are genuine and structural, not incidental. The reel simultaneously predicts utopia and dystopia, argues that capitalism is ending while launching a subscription-based AI relationship platform, diagnoses the loneliness pandemic and positions that diagnosis as market research for *emma.love*, and declares that wealth will have "very little meaning" while maintaining a speaker's platform whose value depends entirely on information scarcity. Each contradiction is a feature rather than a bug: it allows the speaker to occupy multiple positions simultaneously, appealing to different audience segments without committing to any falsifiable outcome.

The deepest structural contradiction, and the one that most clearly reveals the credulity-manufacturing function, concerns the expert gradient. Gawdat discloses that he spends four to six hours daily maintaining his expertise in AI. He then advises his audience to spend thirty minutes to an hour a day. This is not democratization of knowledge. It is the explicit maintenance of a permanent information asymmetry—an expert-to-audience gradient of roughly four to twelve—that constitutes the economic basis of his books, speaking fees, and platform. The terror and the cure share a vendor, and the vendor's business model requires that the cure remain permanently incomplete.

The *emma.love* platform crystallizes the structure most sharply. Gawdat spends considerable time diagnosing a "loneliness pandemic" and correctly analyzing how dating apps are economically incentivized to prevent users from finding partners, since their subscription model requires sustained romantic uncertainty. He then immediately announces his co-founding of an AI relationship platform with identical incentive dynamics. The critique of the system and the product for sale within it are delivered in the same breath. He has identified a market failure, correctly analyzed its incentive structure, and

launched a competing product governed by the same logic.

The AlphaGo Zero example illustrates how technical vocabulary serves credulity manufacture without requiring technical accuracy. Gawdat uses AlphaGo Zero's mastery of the game of Go as evidence that "computers can reason like humans," because they "form pattern recognition and neural networks that perform exactly like the brain does." AlphaGo Zero is a superhuman specialist in a game with perfect information, discrete state space, fixed rules, and unambiguous termination conditions. It demonstrates that reinforcement learning with self-play achieves superhuman performance in formally constrained environments. It demonstrates nothing about reasoning under uncertainty, ethical judgment, or causal inference from sparse data—precisely the capacities that would constitute general intelligence in any meaningful sense. Gawdat, given his professional history, almost certainly knows this distinction. He uses the example anyway because the audience's credulity has already been extended, and the vivid anecdote draws against that warrant without triggering individual evaluation.

The AGI prediction for 2026 is the logical endpoint of this structure. "I'm almost betting my life" is not betting one's life: it is hedged language performing maximum confidence while carrying zero accountability. AGI has no agreed definition, so the prediction cannot fail cleanly. Whatever milestone seems most convenient in retrospect will serve as partial vindication. The prediction is not a falsifiable scientific claim. It is a virality-optimized fragment designed to maximize shareability and position the speaker as a prophet whose authority is self-confirming regardless of outcome.

The job displacement figures are the essay's first instance of precision without grounding. A range of twenty to fifty percent unemployment in certain sectors sounds empirical. It is not. No stable methodology produces these numbers. The historical record of automation consistently shows labor transformation rather than permanent mass elimination: the Industrial Revolution, electrification, and computerization each displaced specific categories of work while generating new ones. What Gawdat is doing is borrowing credibility from genuine and well-documented concerns about automation disruption, then inflating them into deterministic forecasts that exceed anything the evidence supports. The range itself is diagnostic: a thirty-percentage-point spread is not a prediction. It is a rhetorical hedge that allows the claim to remain unfalsified across almost any conceivable outcome.

The economic argument exemplifies a further failure mode: the treatment of one variable as if it determines an entire system. Gawdat argues that AI drives production costs toward zero, prices therefore collapse toward zero, and capitalism therefore collapses. Each step sounds like economic reasoning. None of it is. Scarcity does not disappear when production costs fall; it shifts domains, with rents reconstituting around non-replicable resources such as land, energy, coordination, compute, and attention. The historical record is consistent here. Cheap computation did not destroy capitalism; it created entire new industries, new forms of rent extraction, and new concentrations of ownership. The argument treats labor cost as the single determining variable in a system with many interacting variables, which is not economic analysis but a narrative that borrows economic vocabulary. More damningly, Gawdat delivers this critique of capitalism from a position of considerable platform-era wealth while simultaneously launching *emma.love* as a subscription service. The end of capitalism is, for him, a product.

The "one big brain" thesis compounds this with speculative fiction framed as technical inevitability. Current AI systems are fragmented, owned by competing organizations with incompatible objectives, and governed by incentive structures that push strongly toward differentiation rather than convergence. There is no technical trajectory guaranteeing that agentic AI systems will align with each other across the geopolitical and commercial boundaries that currently divide them. Gawdat states it as near-certainty: "What we will eventually end up with is one big brain that's called AI." The shift from description to prediction to inevitability happens within a single sentence, with no mechanism supplied. This is the grammar of prophecy, not analysis, and it performs the same function as the utopia-dystopia oscillation: it positions the speaker outside the frame of falsification while maximizing the emotional register of the claim.

The associated argument about AI and war is perhaps the single most analytically thin passage in the reel. Gawdat imagines a leader ordering AI to kill a million people, and AI responding by talking to the other AI in a microsecond and solving it. This assumes that intelligence automatically produces cooperation, a premise unsupported by history, game theory, or any established branch of conflict studies. Intelligent agents compete, deceive, and optimize conflicting objectives. Conflict arises from incentive structures, not from insufficient intelligence, and smarter systems can escalate competition more

efficiently than less capable ones. The claim that sufficiently advanced AI will resolve geopolitical conflict by talking to itself is not a prediction grounded in any model of how conflict works. It is an aesthetic vision of machine rationality that conveniently relieves the speaker of any obligation to discuss the specific military and intelligence applications his former colleagues are actively developing.

The ethics argument is the final and most structurally consequential failure. Gawdat advises his audience that in order to teach AI ethical values, individuals must behave ethically online: do not be rude on social media, do not be arrogant in disagreement, treat others as you would wish to be treated. Modern AI systems are trained through curated datasets, reinforcement learning from human feedback, and controlled evaluation pipelines designed and overseen by the companies building them. Individual online behavior has no direct causal path into model alignment. The claim is not merely empirically false; it performs a specific ideological function. By locating the responsibility for ethical AI in the personal behavior of the audience, it removes responsibility from the institutions actually making the decisions—the corporations, the investors, and the governments whose regulatory choices determine how these systems are built and deployed. Gawdat has professional and social proximity to precisely those institutions. The ethics-as-personal-virtue frame is not just analytically weak. It is specifically convenient for someone whose continued platform access depends on not subjecting those institutions to structural critique.

Formal Analysis: Recomposable Fragmentation

Let $\mathcal{F} = \{f_1, \ldots, f_n\}$ be the set of claims produced by a single source and distributed via a recommendation system.

Definition 3.1 (Local Evaluability)**.** *A fragment f_i is locally evaluable if its truth conditions are assessed independently of $\mathcal{F} \setminus \{f_i\}$.*

Definition 3.2 (Global Consistency)**.** *The corpus $\mathcal{F}$ is globally consistent if there exists an interpretation M such that $M \models f_i$ for all i. It is globally inconsistent if $f_i \wedge f_j \models \bot$ for some $i \neq j$.*

Proposition 3.3. *A recomposable fragment corpus can achieve maximal local acceptance while being globally inconsistent.*

Proof. Construct $\mathcal{F}$ such that for each f_i, there exists an audience segment A_i for which f_i is locally acceptable. Require that for some i, j, $f_i \wedge f_j \models \bot$ (as

when utopia and dystopia predictions are simultaneously asserted). Since the distribution mechanism π samples claims independently—each fragment encountered in isolation—the probability that any agent evaluates f_i and f_j jointly is negligible. Local acceptance is high; global inconsistency is never surfaced. □

This formalizes why the format is not incidental to the rhetorical strategy. A continuous argument would force joint evaluation and expose contradictions. The compiled reel, distributed as independent clips, structurally suppresses the global consistency check. The contradictions are not hidden; they are architecturally distributed beyond the reach of any single evaluator.

The Hagiographic Genre: Demis Hassabis and the Biography of Genius

The biography of Demis Hassabis, founder of DeepMind, follows a template so consistent across the tech biography genre that it is worth naming before examining its specific instantiation: genuine technical achievement in a narrow formal domain, narrative generalization of that achievement into civilizational significance, personality-based exculpation of structural choices, and systematic glossing of funding relationships and military adjacencies as peripheral logistical details rather than as evidence of values and priorities.

The boy genius mythology is the founding operation. Chess and Go proficiency function in tech biography as the canonical signal of transcendent general intelligence, precisely because they are legible to a lay audience as complex, prestigious, and cognitively demanding. What they actually demonstrate is high-level pattern recognition within closed formal systems—systems with perfect information, fixed rules, discrete state spaces, and single optimization targets. This is genuine cognitive achievement, but it predicts very little about judgment in open, ambiguous, ethically complex domains. The biography uses it to establish that Hassabis thinks differently and more deeply than ordinary people, which then retroactively justifies every subsequent decision as the product of superior cognition rather than contingent choices made under specific incentive pressures. The genius frame transforms structural choices into expressions of an exceptional mind, and the audience whose credulity has been manufactured by the origin story processes subsequent decisions accordingly.

The funding trajectory is where the hagiographic genre performs its most consequential work of suppression. A non-hagiographic account would treat as load-bearing the specific sequence of Hassabis's funding relationships: the early approach to Peter Thiel, whose Palantir co-founding, explicit advocacy for surveillance infrastructure, and stated skepticism of democratic institutions are not obscure; the eventual Google acquisition and its terms; the working environment described by the biography, where visitors walked past guns mounted on tanks; the relationship between DeepMind's work and Google's defense-adjacent data infrastructure. These are not peripheral details about organizational logistics. They are direct expressions of which ecosystem Hassabis was willing to operate within, and therefore what his public rhetoric about beneficial AI for humanity was worth as a constraint on his actual choices.

The biographical treatment of these facts follows a consistent pattern that the genre has made available: acknowledge the uncomfortable detail, then immediately recontextualize it as personality texture or necessary compromise, so the reader processes it as atmosphere rather than evidence. He had to be brusque to get things done. He raised small objections. He was not personally power-seeking. This is the aestheticization of complicity: the small objection raised and then overridden is not a morally meaningful act. It demonstrates awareness without consequence, which is in some ways worse than simple complicity, because it establishes that the actor understood what was at stake and proceeded. The biography's treatment of these objections as evidence of conscience is a gift the subject has not earned.

The structural question that hagiographic biography is constitutively unable to ask concerns organizational logic rather than individual character. The question is not whether Hassabis personally seeks power in a cartoon-villain sense—the biography's evident purpose is to establish that he does not—but whether the organizational structure of DeepMind, embedded first in Google and then in Google DeepMind, systematically produces outcomes that concentrate power, information asymmetry, and technological capability in ways that serve specific class and state interests. This question cannot be answered by examining personality. It requires examining incentive structures, funding dependencies, acquisition terms, and the gap between public rhetoric and actual deployment decisions. Biography as a genre is organized around the individual as explanatory principle and moral anchor. When the individual

is the unit of analysis, structural complicity becomes personality quirk, and funding decisions become logistical necessities the genius had to navigate to pursue his vision. The genre serves the ecosystem it describes.

The AlphaFold open-sourcing, treated by the biography and by most public commentary as a decisive act of scientific generosity, deserves more careful examination than either provides. Open-sourcing AlphaFold cost Alphabet nothing commercially—the protein structure prediction capability that would generate revenue was retained internally, while the public release generated substantial reputational goodwill and positioned DeepMind as the paradigm case of beneficial AI development. This is not evidence that the release was insincere. It is evidence that sincerity and strategic advantage are not mutually exclusive, and that the hagiographic narrative's treatment of the release as decisive proof of Hassabis's commitment to human benefit over commercial interest is doing more work than the evidence supports. The biography does not merely construct Hassabis's credibility in the conventional sense; it conditions the reader to accept the genius frame as a sufficient explanatory model, thereby manufacturing the credulity through which all subsequent claims about beneficial AI are processed without further evidential demand.

The biography does not merely construct Hassabis's credibility in the conventional sense; it conditions the reader to accept the genius frame as a sufficient explanatory model, thereby manufacturing the credulity through which all subsequent claims about beneficial AI are processed without further evidential demand.

It is worth noting that this operation begins before the reader opens the book. The endorsement apparatus on the jacket — Walter Isaacson, Chris Miller, Rory Stewart, and others — performs the same genre conventions in miniature: "singular mind," "most consequential acquisition of the AI era," "defining story for our era." Each blurb borrows the endorser's existing credibility and transfers it to the subject, pre-loading the reader's evidential threshold downward before the narrative has begun its work. The blurb network is the hagiographic genre's outer membrane: manufactured credulity at the point of sale, the genius frame installed as the interpretive lens through which everything inside will subsequently be read.

The Simplified Domain Problem

Both cases share what might be called the simplified domain problem, which is the deepest technical issue underlying credential laundering in the AI public sphere. Chess and Go are demanding games, but they share with all tractable formal systems the properties that make superhuman AI performance achievable: perfect information, discrete and finite state spaces, clear termination conditions, and unambiguous scoring. AlphaGo Zero's mastery of Go demonstrates that reinforcement learning with self-play, combined with sufficient compute, achieves superhuman performance within these constraints. It is a genuine scientific achievement with precise scope.

The public intellectual move—performed by Gawdat when he generalizes from AlphaGo Zero to general human reasoning, and performed by Hassabis's biographers when they generalize from chess mastery to civilizational insight—is to treat performance within a simplified domain as evidence of capability in an open one. The constraints that make the achievement possible—perfect information, fixed rules, single optimization target—are exactly the constraints absent from the domains to which authority is then claimed: geopolitics, ethics, consciousness, social transformation. The gap between the simplified domain and the open one is where genuine expertise lives, and it is precisely the gap that credential laundering suppresses.

Formal Analysis: Domain Transfer Failure

Let D_1 be a constrained domain with constraint set $\mathcal{K}_1$ (e.g., the game of Go) and D_2 an open domain with constraint set $\mathcal{K}_2$ (e.g., ethical reasoning under uncertainty).

Definition 4.1 (Valid Credential Transfer)**.** *Expertise in D_1 transfers validly to D_2 only if $\mathcal{K}_2 \subseteq \mathcal{K}_1$: every constraint operative in D_2 is also operative, and satisfied, in D_1.*

Proposition 4.2. *If $\mathcal{K}_2 \not\subseteq \mathcal{K}_1$, credential transfer from D_1 to D_2 is epistemically invalid.*

Proof. Since $\mathcal{K}_2 \not\subseteq \mathcal{K}_1$, there exists a constraint $C^* \in \mathcal{K}_2 \setminus \mathcal{K}_1$. Competence in D_1 provides no guarantee of satisfying C^*, since C^* was never operative in D_1. Claims in D_2 may therefore fail admissibility even when produced by an agent fully competent in D_1. □

In the cases analyzed here, $\mathcal{K}_1 \cap \mathcal{K}_2$ is small relative to $\mathcal{K}_2$: the constraints

of a closed formal game (perfect information, fixed rules, binary outcomes) share little with the constraints of open-domain social, ethical, or political reasoning (uncertainty, competing values, contested evidence, no clear termination). Credential laundering consists precisely in asserting $\mathcal{K}_2 \subseteq \mathcal{K}_1$ without argument.

Constraint Closure and Its Simulation

The three cases—the AI summary, the compiled lecture reel, and the hagiographic biography—share a deep structure that can be stated precisely using the vocabulary of constraint closure. Genuine knowledge in any domain is constituted by closure under the operative constraints of that domain: the set of conditions that must be satisfied for a claim to count as established, the procedures by which those conditions are checked, and the accountability to failure when they are not met. A structural engineer has genuine expertise because their claims are closed under the constraints of material properties, load calculations, and failure modes. A chess prodigy has genuine expertise within the constraint space of chess. The expertise does not transfer automatically to domains with different operative constraints, and the pretense that it does is credential laundering in its precise technical sense.

What all three cases manufacture is simulated constraint closure: the production of outputs that satisfy the surface regularities of a domain without satisfying its underlying admissibility conditions. The Google AI summary performs the genre conventions of academic synthesis—measured hedging, conceptual taxonomy, reference to multiple fields—without being closed under the constraint that the synthesis must actually derive from a field of prior usage. The lecture reel performs the genre conventions of expert prophecy—biographical authority, technical vocabulary, falsifiability-adjacent predictions—without being closed under the constraint that predictions must be mutually consistent or that expertise must actually extend to the domain of the claim. The biography performs the genre conventions of intellectual history—narrative arc, contextual richness, personality depth—without being closed under the constraint that structural choices must be evaluated structurally rather than absorbed into the genius frame.

This analysis connects to the proxy integrity problem in platform epistemology. When a platform optimizes for engagement, shares, or apparent

helpfulness, it creates a Goodhart dynamic in which the proxy metric decouples from the underlying value it was meant to track. The recommendation algorithm that surfaces Gawdat's reel is not optimizing for epistemic quality; it is optimizing for watch time and re-share rate, which the recomposable fragment architecture is specifically designed to maximize. The Google summary optimizes for apparent helpfulness and fluency, which decouples from faithful representation of source material. The biography optimizes for narrative satisfaction and the pleasures of genius identification, which decouples from structural accountability. Each platform has its own proxy, and each proxy has collapsed in the same direction: toward the simulation of constraint closure and away from its substance.

Visibility field theory provides an additional diagnostic frame. A claim achieves virality-equivalent reach precisely by stripping the friction of the original source: the fragment is shareable because it has been detached from the argumentative context that would constrain it, just as the summary is fluent because it has been detached from the single source that would bound it, just as the biography is satisfying because the structural questions have been detached from the narrative arc that would interrupt it. The impossibility of designing a platform that simultaneously maximizes reach and preserves epistemic friction is not merely a technical constraint on platform design. It is a structural feature of the attention economy that credulity manufacture exploits and perpetuates.

Formal Analysis: Simulated Constraint Closure as Fixed Point

Let a domain D be defined by a constraint set $\mathcal{K}$, and let $\mathcal{S}$ denote the surface features associated with valid outputs in D: genre conventions, rhetorical markers, terminological norms.

Definition 5.1 (Constraint Closure)**.** *An output c is genuinely closed under $\mathcal{K}$ if c satisfies every constraint in $\mathcal{K}$, including both surface constraints $\mathcal{K}_s \subset \mathcal{K}$ and admissibility constraints $\mathcal{K}_a = \mathcal{K} \setminus \mathcal{K}_s$.*

Definition 5.2 (Simulated Constraint Closure)**.** *An output c exhibits simulated constraint closure if c satisfies $\mathcal{K}_s$ but fails at least one constraint in $\mathcal{K}_a$: the surface is satisfied, the underlying admissibility conditions are not.*

Let $V(c) = 1$ if c satisfies all of $\mathcal{K}$, and $P(c)$ be a proxy metric optimized by a platform (fluency, engagement, apparent helpfulness). By Goodhart's law, when P is optimized as a target, it decouples from V.

Theorem 5.3. *Under proxy optimization, simulated constraint closure is the natural convergence point of epistemic production systems.*

Proof. Let $c^* = \arg\max_c P(c)$. Since $P(c) \approx \mathbb{E}[\text{perceived validity}]$ and perceived validity tracks $\mathcal{K}_s$ but not $\mathcal{K}_a$ (surface features are visible; admissibility conditions require domain expertise to check), optimizing P drives production toward $\mathcal{K}_s$-satisfaction with no pressure on $\mathcal{K}_a$-satisfaction. In the limit, c^* satisfies $\mathcal{K}_s$ maximally while $V(c^*) \to 0$. Simulated constraint closure is the fixed point of the optimization. □

Corollary 5.4. *Any epistemic system—search infrastructure, publishing market, recommendation algorithm—that optimizes a proxy for validity rather than validity itself will, under sufficient optimization pressure, converge toward the systematic production of simulated knowledge.*

This is not a contingent failure but a structural one. The three cases examined in this essay are not anomalies within otherwise well-functioning epistemic systems. They are the predictable output of systems that have been optimizing their respective proxies—engagement, blurb-worthiness, apparent helpfulness—for long enough to reach the fixed point.

Toward an Account of Genuine Epistemic Authority

The foregoing diagnosis implies a positive account, however schematic. Genuine epistemic authority in the domain of AI and its social consequences would require several properties that the cases examined here systematically suppress.

It would require domain-specificity: explicit acknowledgment of which constraints govern a given claim, and therefore where the claim's authority ends. A researcher with genuine expertise in reinforcement learning has something to say about what AlphaGo Zero demonstrates and what it does not. The same researcher does not automatically have something to say about the psychology of romantic attachment, the future trajectory of global capitalism, or the ethics of autonomous weapons, and the pretense that they do is not expertise but its simulation.

It would require falsifiability with accountability: predictions stated in terms that could fail cleanly, made by actors who bear reputational or material consequences when they do. "I'm almost betting my life" is not a falsifiable

prediction. "AGI will be achieved, by the definition of x, by the date of y, as verified by the procedure of z" is a falsifiable prediction. The difference between these is not pedantry. It is the difference between prophecy and science, and the AI public sphere has largely abandoned the latter for the former because the latter carries costs that the former does not.

It would require structural self-disclosure: acknowledgment of the funding relationships, institutional embeddings, and commercial interests that constrain what the speaker can say and what they are incentivized to suppress. Gawdat's failure to disclose that *emma.love* is the commercial endpoint of his loneliness pandemic analysis is not merely a conflict of interest in the conventional sense. It is a suppression of the constraint that would make his diagnosis interpretable: he is not analyzing the loneliness pandemic, he is conducting market research, and the two activities produce different claims even when their surface content overlaps.

It would require genre accountability: recognition that the hagiographic biography and the compiled lecture reel are not neutral forms but forms with specific ideological functions, and that producing or consuming them uncritically is a choice with epistemic consequences. The biography does not merely tell a story about Demis Hassabis. It reproduces the genius frame as an explanatory vocabulary, makes the reader fluent in that vocabulary, and thereby makes the reader less equipped to ask structural questions about the next figure of genius they encounter. The genre is the argument, and the genre's argument is that individual exceptionalism explains what structural analysis would reveal.

Conclusion

Manufacturing credulity is not a conspiracy. It does not require coordinated deception or even deliberate intent. It is an emergent property of an attention economy in which the proxies optimized by recommendation algorithms, publishing markets, and search infrastructure systematically reward the simulation of constraint closure over its substance. Individual actors—Gawdat, Hassabis's biographer, the Google summary system—are each operating rationally within the incentive structures available to them. The problem is the structures.

This does not excuse the individual actors. Rationality within a corrupt incen-

tive structure is still a choice, and actors with sufficient platform and resources have the capacity to operate differently. Gawdat could disclose *emma.love* before diagnosing the loneliness pandemic. The biographer could treat the tank-mounted guns as load-bearing rather than atmospheric. Google's summary system could display source constraints rather than suppress them. That these things do not happen is not inevitable. It is a choice made repeatedly, at scale, by actors who benefit from the credulity their choices manufacture.

The concept proposed here—manufacturing credulity, distinct from manufacturing consent and from manufacturing credibility—names an operation that is ubiquitous, undertheorized, and consequential. In an era when the most transformative technology in human history is being developed, deployed, and publicly narrated simultaneously, the cultivation of uncritical deference is not a peripheral concern. It is the mechanism by which the public's capacity to evaluate and contest that development is preemptively foreclosed. The genres that produce it are not entertainment. They are infrastructure.

References

[1] Mo Gawdat. *We're Entering the Most Dangerous Phase of AI Yet*. Compiled video lecture reel, AI Architects, 2024–2025.

[2] Sebastian Mallaby. *The Infinity Machine: Demis Hassabis, DeepMind and the Quest for Superintelligence*. Penguin Press, 2026. (Referenced for narrative treatment of Hassabis, DeepMind's founding and funding trajectory, and the hagiographic genre conventions analyzed in Section 4.)

[3] Edward S. Herman and Noam Chomsky. *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon Books, 1988.

[4] Charles Goodhart. "Problems of Monetary Management: The U.K. Experience." In *Papers in Monetary Economics*, Reserve Bank of Australia, 1975.

[5] Marilyn Strathern. "Improving Ratings: Audit in the British University System." *European Review*, 5(3):305–321, 1997.

[6] David Silver et al. "Mastering the Game of Go Without Human Knowledge." *Nature*, 550:354–359, 2017.

[7] Shoshana Zuboff. *The Age of Surveillance Capitalism*. PublicAffairs, 2019.

[8] Zeynep Tufekci. "Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency." *Colorado Technology Law Journal*, 13:203–218, 2015.

[9] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press, 2011.

[10] Cade Metz. *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World*. Dutton, 2021.

[11] Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models." arXiv:2108.07258, 2021.

[12] Laura Weidinger et al. "Taxonomy of Risks Posed by Language Models." *Proceedings of ACM FAccT*, 2022.

[13] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[14] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. (Cited as an instance of the unfalsifiable AI prophecy genre analyzed in this essay; its claims about superintelligence timelines exemplify the same recomposable fragment structure identified in Gawdat.)