Artificial Intelligence Is a Misnomer: Weird Quotes, Metaphors, and the Politics of Naming AI

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

January 27, 2026

“Weird quotes about AI” are not just internet seasoning. They are compressed arguments—miniature philosophies—that try to do three things at once: explain unfamiliar systems, warn about misuse, and compete for narrative control. When someone says “AI is just curve fitting,” “a blurry JPEG of the web,” “alchemy,” “a stochastic parrot,” or “not intelligence but cleverness,” they are making a claim about what the technology is, what it is not, and what kinds of trust it deserves. These lines spread because they are memorable, but that same memorability can detach them from context and turn them into slogans. This paper treats weird quotes as a serious dataset: rhetorical artifacts that reveal fault lines between engineering reality and public imagination, between philosophical theories of mind and practical product behavior, and between governance needs and marketing incentives. We map recurring metaphors, identify what each one illuminates and what it hides, and propose a reader’s framework for evaluating quotable claims without becoming either credulous or cynical. The goal is not to “debunk quotes,” but to use them to clarify the real stakes: understanding, agency, responsibility, and how language itself steers the future of AI.

Keywords: artificial intelligence, large language models, machine learning, AI hype, AI skepticism, AI metaphors, naming and framing, artificial cleverness, stochastic parrots, blurry JPEG of the web, curve fitting, algorithmic alchemy, automation bias, anthropomorphism, interpretability, philosophy of mind, media ecology, science communication, AI governance, risk narratives.

A collaboration with GPT-5.2-Thinking. 55 pages.

Outline

1. Introduction: Why Weird Quotes About AI Matter
1.1 Quotes as compressed theories, not jokes
1.2 The “quote laundering” problem: context loss and slogan drift
1.3 What this paper will and will not do

2. What Counts as a Weird Quote
2.1 Misnomers, metaphors, inversions, and provocations
2.2 When a quote is descriptive vs prescriptive
2.3 The importance of original context: audience, date, and purpose

3. A Taxonomy of Weirdness: The Main Families of AI One-Liners
3.1 Misnomer quotes: “not intelligence, just cleverness”
3.2 Compression quotes: “lossy summaries” and “blurry JPEG” framings
3.3 Mimicry quotes: parrots, plagiarism, and pattern replay
3.4 Magic quotes: alchemy, sorcery, and “black box” mystique
3.5 Monster quotes: demons, existential dread, runaway agency
3.6 Tool quotes: “augmentation,” “adjunct,” “calculator,” “prosthesis”
3.7 Deflation quotes: “just statistics,” “just autocomplete,” “just curve fitting”
3.8 Inversion quotes: “the stupidity of AI” and anti-genius framings

4. Case Studies: What the Famous Weird Quotes Get Right and Wrong
4.1 “Artificial cleverness”: misnomer as a theory of understanding
4.2 “Neither artificial nor intelligent”: materials, labor, and ideology
4.3 “Stochastic parrots”: fluency, grounding, and social risk
4.4 “Blurry JPEG of the web”: compression, generalization, and failure modes
4.5 “Alchemy”: engineering without theory and the limits of interpretability
4.6 “Curve fitting”: representation, causality, and where it becomes misleading
4.7 “Autocomplete on steroids”: UI metaphors and underclaim/overclaim traps
4.8 “AI effect”: why working systems stop being called AI

5. Why Smart People Speak This Way
5.1 Communication constraints: interviews, headlines, and virality
5.2 Incentives: funding, warning, branding, and reputational defense
5.3 Philosophical priors: mind, meaning, consciousness, and computation
5.4 The pedagogical motive: metaphors as scaffolding for non-experts

6. What Metaphors Illuminate and What They Hide
6.1 What each family highlights: capability, risk, or epistemic limits

1. Introduction: Why Weird Quotes About AI Matter

Walk into any serious conversation about modern AI and you will eventually hear a one-liner that seems to arrive from a different genre than engineering: “AI is just curve fitting,” “a blurry JPEG of the web,” “stochastic parrots,” “alchemy,” “autocomplete on steroids,” “neither artificial nor intelligent,” “not intelligence but cleverness.” These lines sound like jokes, provocations, or rhetorical flourishes. But they persist because they are doing real work. They compress a worldview about what AI systems are, what they are not, what kinds of trust they merit, and what risks society should prioritize.

In public life, naming is power. Labels determine what gets regulated, funded, celebrated, feared, or normalized. “Artificial intelligence” is not a neutral technical descriptor; it is a story with implied agency, implied competence, and implied inevitability. Weird quotes are the immune response to that story. They are attempts to puncture hype, correct category errors, or shift the debate from metaphysics to accountability. They also sometimes function as their own kind of hype, by lending drama and destiny to what might otherwise be understood as an evolving stack of statistical methods, data pipelines, and product interfaces.

This paper treats weird quotes as an evidence stream: a cultural dataset that reveals recurring confusions and recurring truths, and that helps us map where experts, critics, industry, journalists, and the public are talking past each other. The aim is not to mock the quotes or worship them, but to use them as diagnostic instruments for the present moment.

1.1 Quotes as compressed theories, not jokes

A good one-liner is not merely a witty soundbite; it is a compressed theory of mind, language, and technology. When someone says “AI doesn’t understand,” they are implicitly asserting a definition of understanding. When someone says “it’s just statistics,” they are implicitly asserting which aspects of intelligence matter and which do not. When someone calls ML “alchemy,” they are making a claim about epistemology: the gap between what works in practice and what is explained in principle. When someone calls AI “a tool, not a creature,” they are making a claim about moral and legal responsibility: tools have owners and operators; creatures have intentions and rights.

Compression is not a flaw; it is a feature of human communication under constraint. Most audiences will not read a technical paper, a model card, or a careful philosophical argument. They will, however, remember a vivid metaphor. In that sense, weird quotes are a practical technology of meaning: they enable non-experts to carry a mental model of something too complex to fully parse. They can be pedagogical scaffolding.

But compressed theories always trade accuracy for portability. A metaphor highlights one dimension while erasing others. “Blurry JPEG” highlights lossy compression and artifacts; it

hides the ways models can generalize beyond direct copying. “Autocomplete” highlights surface-level next-token prediction; it hides emergent planning behaviors, tool use, and multi-step competence under scaffolding. “Curve fitting” highlights the reliance on statistical pattern extraction; it hides representation learning and the ways learned internal structures can support nontrivial transfer.

The deeper point is that each quote tends to aim at a particular target: anthropomorphism, mystification, marketing, fear, or complacency. Weird quotes are not random; they are rhetorical countermeasures. They recur because the same misconceptions recur: that fluency equals truth, that confidence equals competence, that a system with language must have a mind, that automation implies objectivity, that “intelligence” is a single substance rather than a family of capabilities and dispositions.

If we treat these quotes as tiny arguments rather than entertainment, we can ask better questions: What definition of intelligence is being used? What failure mode is being warned against? What hidden assumption about agency is being smuggled in? What policy preference follows if we accept the framing?

1.2 The “quote laundering” problem: context loss and slogan drift

The modern media ecosystem rewards lines that can travel. A nuanced claim made in a long interview gets clipped into a fifteen-second segment. A careful metaphor becomes a headline. A headline becomes a meme. The meme becomes a talking point. At each step, context falls away, and the quote gains rhetorical flexibility. This is quote laundering: the process by which an utterance becomes more portable and less accountable.

The laundering mechanism has several predictable stages. First, selection: the most vivid or polarizing sentence is extracted, not the most representative. Second, compression: surrounding qualifiers and definitions are removed. Third, retargeting: the quote is applied to a broader or different claim than the speaker intended. Fourth, authority transfer: the speaker’s prestige becomes a stamp that appears to certify whatever the quote is now being used to argue.

Slogan drift is the downstream effect. A phrase coined to critique overclaiming in one domain gets applied as a blanket dismissal of all progress. A warning about lack of grounding becomes an argument that the technology is useless. Or the opposite: a dramatic metaphor intended as caution is recycled as proof of inevitability and power, feeding the very hype it was meant to resist.

Quote laundering matters because it changes decision-making. In institutions, slogans become proxies for evaluation. Teams adopt or reject systems based on vibes: “it’s just autocomplete” versus “it’s basically a new species.” Regulators face pressure to legislate with incomplete

technical vocabulary, and portable quotes become substitute definitions. Even within AI research and engineering, memorable framings can bias what gets built, measured, or funded. A quote can become a shortcut that bypasses actual model behavior and actual deployment outcomes.

This paper takes laundering seriously as a phenomenon to analyze, not simply to complain about. The goal is to reconstruct context where possible, to identify what claims survive decontextualization, and to develop tools for readers to notice when a quote is being used as evidence versus used as a weapon.

1.3 What this paper will and will not do

This paper will do five things. First, it will treat weird quotes as a structured dataset rather than a list of curiosities. Second, it will build a taxonomy of the major families of AI one-liners, organized by the kind of claim they implicitly make: misnomer claims, compression claims, mimicry claims, magic claims, monster claims, tool claims, deflation claims, and inversion claims. Third, it will unpack what each family tends to illuminate and what it tends to hide, so readers can see both the insight and the blind spot. Fourth, it will offer a framework for evaluating quotable claims: context tests, mechanism tests, scope tests, and stakes tests. Fifth, it will show how these quotes shape governance and public understanding by functioning as informal definitions in policy, procurement, and media narratives.

This paper will not attempt to “settle” whether machines can ever truly understand, whether consciousness is computable, or whether intelligence is fundamentally symbolic, statistical, embodied, or spiritual. Those debates are important, but they are not the most urgent bottleneck for most real-world decisions being made about AI systems right now. It will also not compile an exhaustive anthology. The point is not completeness; the point is pattern recognition. Nor will it treat every quote as equally serious. Some are careful distillations of a position; others are rhetorical grenades. Distinguishing between them is part of the task.

Finally, the paper will not assume that weird quotes are inherently anti-AI. Many are pro-clarity rather than anti-capability. They can be used to demand better evaluation, better governance, and better language that matches what systems actually do. In an environment where both hype and backlash can be irrational, the discipline is the same: translate slogans back into testable claims and concrete responsibilities.

2. What Counts as a Weird Quote

A “weird quote about AI” is not merely a funny line or a harsh takedown. It is a compact reframing that collides with the listener’s default picture of AI. The weirdness comes from

mismatch: the quote refuses the polite story (AI as mind, AI as genius, AI as neutral tool) and substitutes a different model that is often more vivid, more moral, or more deflationary. In practice, weird AI quotes function like cognitive pry bars. They force a shift in categories: from "intelligence" to "cleverness," from "knowledge" to "compression," from "agency" to "automation," from "science" to "alchemy," from "tool" to "demon," from "smart" to "stupid." Whether that shift clarifies or distorts depends on what the quote is trying to do, and on the context it came from.

For the purposes of this paper, a weird quote has three traits. First, it is memorable enough to travel. Second, it makes an implicit claim about the nature of AI systems or the social meaning of AI deployment. Third, it tends to be used as a proxy for a broader argument. A quote that is merely insulting or merely technical is usually not "weird" in this sense. Weirdness is the signature of a metaphor doing argumentative work.

2.1 Misnomers, metaphors, inversions, and provocations

We can group weird AI quotes into four overlapping modes. These are not rigid categories; they are rhetorical strategies that often blend.

Misnomers are attempts to rename the thing. They claim the headline term is fundamentally misleading. "Artificial intelligence is a misnomer" is a classic example, but the deeper move is to reject what the word "intelligence" smuggles in: understanding, intention, agency, wisdom, moral competence. Misnomer quotes often propose a substitute label such as "cleverness," "automation," "statistics," "pattern recognition," or "prediction." Their ambition is definitional. If you control the definition, you control the debate. You can decide what counts as success, what counts as risk, and what must be regulated.

Metaphors are attempts to explain the thing by mapping it onto a familiar domain. "Blurry JPEG of the web" is a metaphor of lossy compression. "Stochastic parrots" is a metaphor of mimicry and repetition. "Autocomplete on steroids" is a metaphor of incremental continuation with an amped-up scale. "Alchemy" is a metaphor of practice without theory, craft without explanation, recipes without guarantees. Metaphors are powerful because they feel like understanding even when they are only partial. They give the mind a handle. They can be pedagogical, but they can also become ideological: once a metaphor takes hold, it is hard to see what it hides.

Inversions flip the expected polarity. Instead of "superintelligence," you get "the stupidity of AI." Instead of "AI is objective," you get "AI is neither artificial nor intelligent." Instead of "AI reasons," you get "AI hallucinates." Inversions are effective because they exploit surprise. They are a rhetorical judo move: they use the dominant narrative's momentum against itself. But inversions can also overcorrect. They risk turning a legitimate critique of boundaries into a blanket denial of capability.

Provocations are lines designed to start a fight, not to teach. “We are summoning the demon” is not a technical description; it is a moral alarm bell. Provocations often show up when speakers believe normal language cannot produce sufficient caution, or when they believe the public is being lulled by marketing. Provocations can be ethically motivated, but they are also media-adapted: they travel better than caveats. They can spur real scrutiny, or they can polarize audiences into camps that stop listening to each other.

What matters is that all four modes are attempts to reset the frame. They are not neutral descriptions of a machine. They are interventions into a narrative war over what AI is allowed to mean.

2.2 When a quote is descriptive vs prescriptive

Weird quotes often slide between describing what AI systems do and prescribing how society should treat them. This is a crucial distinction because it determines what kind of evidence is relevant.

A descriptive quote is, at least in intention, about the system’s behavior or internal structure. “It’s next-token prediction” is descriptive. “It compresses patterns” is descriptive. “It lacks grounding” is descriptive. The proper response to a descriptive quote is empirical: test it, probe it, measure it, ask what conditions make it more or less true. Descriptive quotes can still be biased or simplistic, but they are, in principle, answerable by observation and experiment.

A prescriptive quote is about values, governance, or moral posture. “Don’t let it make decisions” is prescriptive. “Treat it like a tool” is prescriptive. “Regulate it like a hazardous technology” is prescriptive. “Don’t anthropomorphize it” is prescriptive. The proper response here is normative and institutional: who is accountable, what rights are protected, what harms are plausible, what tradeoffs are acceptable, what safeguards are feasible.

Many quotes masquerade as descriptive while sneaking in a prescription. “It’s just autocomplete” is often deployed not merely as a description, but as a dismissal: therefore it should not be trusted, funded, or admired. Conversely, some quotes masquerade as prescriptions while leaning on an unexamined description. “We are summoning a demon” presupposes certain kinds of uncontrollability and agency. If those assumptions are wrong, the prescription may still be prudent, but it needs a different justification.

This paper treats the descriptive-prescriptive distinction as one of the most important reader tools. A quote can be emotionally compelling and still be logically ambiguous about which type it is. When a line is traveling fast, it is often doing both at once. The critical move is to separate the layers: what claim about the system is being made, and what claim about policy or morality is being implied.

2.3 The importance of original context: audience, date, and purpose

Quotes are time-stamped artifacts. A remark made in 2016 about "AI" may have been aimed at a different technological regime than a similar remark made after the rise of frontier-scale language models. A line spoken in a technical workshop functions differently than the same line spoken on a stage designed for viral clips. A quote addressed to policymakers will often emphasize accountability and social consequences. A quote addressed to researchers will often emphasize epistemic limits and scientific standards. A quote addressed to investors or product audiences will often be shaped by incentives to differentiate, warn, or market.

Three context variables matter most: audience, date, and purpose.

Audience determines the expected background knowledge and the rhetorical goals. If the audience is general, the speaker may choose metaphors that are vivid but coarse. If the audience is technical, the speaker may use shorthand that assumes shared definitions. If the audience is hostile, the speaker may sharpen the blade. If the audience is friendly, the speaker may indulge nuance. A quote lifted from one audience and reused in another can become misleading even if it was fair in its original setting.

Date matters because AI systems change rapidly, but also because the surrounding discourse changes rapidly. Some quotes critique a specific wave of hype. Others respond to a specific policy panic. Others answer a question that was live at the time and now has a different shape. When quotes are detached from date, they acquire an illusion of timelessness. That illusion can be exploited. A skeptical line from an era of weaker systems can be weaponized to dismiss new capabilities. A hype line from an era of weaker evaluation can be weaponized to overclaim today.

Purpose is the hardest to see, because it is often mixed. A speaker may be teaching. They may be warning. They may be defending a competing research program. They may be steering a funding agenda. They may be pushing back on a narrative they consider dangerous. They may be trying to avoid liability. They may be trying to cut through noise for a lay audience. A quote is an action inside a conversation. Without the conversation, we misread the action.

This is why "quote laundering" is not just a media problem but an epistemic problem. When a line becomes a free-floating token, it stops being an utterance and becomes an instrument. The reader's discipline is therefore to reconstruct enough of the original scene to know what the quote was doing.

In later sections, this paper will formalize this into a practical method: treat each weird quote as a claim with parameters. Who said it. When. To whom. In response to what question. With what underlying definition of intelligence, understanding, or agency. With what implied prescription.

Only then can a reader evaluate whether the quote is a clarifying lens, a useful warning, an overcorrection, or a rhetorical grenade.

3. A Taxonomy of Weirdness: The Main Families of AI One-Liners

Weird AI quotes are not random fireworks. They cluster into a small number of rhetorical families that reappear because the same underlying tensions reappear: intelligence versus imitation, explanation versus opacity, tool versus agent, awe versus accountability. Each family tends to compress a particular kind of claim about AI systems, and each family carries predictable strengths and predictable distortions. The taxonomy below is not meant to rank quotes as true or false. It is meant to give the reader a map of the argument-space so that when a line shows up in the wild, you can immediately ask: What family is this? What is it trying to highlight? What is it trying to smuggle in? What does it omit?

3.1 Misnomer quotes: “not intelligence, just cleverness”

Misnomer quotes attack the label itself. Their core claim is definitional: the term “Artificial Intelligence” misleads by importing associations that the system does not deserve. The classic move is to replace “intelligence” with something narrower: cleverness, automation, pattern recognition, prediction, statistics, optimization. This family often originates in a moral concern. If you call it “intelligence,” people will grant it authority. They will defer. They will anthropomorphize. They will excuse failures as “mysterious.” They will treat outputs as if they were intentions.

Misnomer quotes tend to highlight an absence: missing understanding, missing meaning, missing consciousness, missing wisdom. They call attention to the gap between human concepts and machine competence. The weirdness is in the refusal: the quote refuses to let language hand the system a crown.

But misnomer quotes can also oversimplify by acting as if “intelligence” is a single property rather than a family of abilities. A system can be non-conscious and still display robust competence across domains. A system can lack understanding in a philosophical sense and still change how institutions operate. Misnomer quotes are powerful for puncturing unjustified deference, but weaker as full technical descriptions.

What this family is best for: resisting anthropomorphic governance and clarifying accountability. What it risks: encouraging dismissal where careful evaluation is needed.

3.2 Compression quotes: “lossy summaries” and “blurry JPEG” framings

Compression quotes treat modern AI, especially language models, as a compression engine for human-produced data. The idea is that the model has ingested enormous corpora and learned a compact representation that can regenerate plausible continuations. “Blurry JPEG of the web” is the emblematic image: you can see the general shape, but fine detail turns into artifacts. The weirdness lies in the demotion of “intelligence” to “compression” and “generation” to “decompression.”

This family is often trying to explain two things at once. First, why models can produce fluent, high-level summaries: compression preserves structure. Second, why models hallucinate and confabulate: compression discards detail and the decompressed output must guess. The metaphor naturally predicts smoothness with occasional grotesque errors, especially at boundaries where the data is sparse, ambiguous, or contradictory.

Compression quotes are pedagogically strong because they connect to everyday experience: images, audio, video, file sizes. They also help correct a common misconception: that fluent text implies stored facts in a reliable database-like form. Compression suggests a different mental model: the model has patterns, not records.

But compression metaphors can mislead when they imply mere reproduction. A strong compression model can generalize; it can combine fragments into new patterns; it can solve tasks that are not simple replay. “It’s just compression” becomes a rhetorical minimization if it ignores the fact that compression and generalization are deeply related in learning theory and in practice.

What this family is best for: explaining hallucination, generalization, and the difference between fluency and truth.
What it risks: collapsing learning into copying and underestimating functional capability.

3.3 Mimicry quotes: parrots, plagiarism, and pattern replay

Mimicry quotes focus on imitation. They suggest the system is essentially a mimic, a parrot, a regurgitator of patterns. The point is not merely that the model trained on human text, but that the model lacks grounding or intent and therefore produces language without understanding. Mimicry metaphors often arrive with ethical claims: the labor and authorship of the training corpus matters, and the model’s output can be seen as derivative appropriation.

The mimicry family highlights a real phenomenon: models can be fluent in contexts they do not “live in.” They can say what a grief counselor might say without grieving; they can speak about war without being wounded. This gap generates both practical risks (false confidence, deception, manipulation) and philosophical questions (meaning, intentionality, reference).

Mimicry quotes also capture an important social hazard: people are easily persuaded by linguistic coherence. A parrot-like metaphor is a warning sign against “authority by eloquence.”

Where mimicry metaphors struggle is in cases where the system demonstrates behavior that looks like more than mere replay: novel combinations, multi-step tool use, cross-domain transfer. Mimicry metaphors can still hold as a critique of grounding and intention, but they become weaker if used to deny all novelty. Also, the plagiarism framing can become a blunt instrument that ignores the spectrum from exact memorization to learned abstraction.

What this family is best for: resisting the seduction of fluency; foregrounding grounding, authorship, and the ethics of training data.
What it risks: treating all generalization as theft or all novelty as illusion.

3.4 Magic quotes: alchemy, sorcery, and “black box” mystique

Magic quotes portray AI as something that works without understanding, like alchemy: a craft of recipes, tinkering, and mysterious effects. “Black box” language fits here, as do metaphors of sorcery and spellcasting. The weirdness is that it places modern computational systems inside a pre-scientific or quasi-mystical frame.

This family usually has two targets. One target is the research culture: an admonition that empirical success has outpaced theoretical explanation, interpretability, and rigorous guarantees. The other target is public mystification: when companies sell models as if they were inscrutable beings rather than engineered artifacts, magic metaphors appear as either critique or reinforcement.

Magic quotes can be clarifying because they spotlight uncertainty, brittleness, and the difficulty of assigning causal explanations to model behavior. They help non-experts understand why two nearly identical prompts can produce wildly different outputs, and why the same model can be impressive and unreliable in close succession.

But magic metaphors can also inadvertently amplify mystique. Calling something sorcery can create a sense of inevitability and helplessness, which pushes institutions toward either fatalism (“we can’t understand it, we just have to accept it”) or panic (“it’s uncontrollable”). In governance terms, magic metaphors can become excuses: “the model did it,” “we don’t know why,” “it’s emergent.” The metaphor that begins as critique can end as abdication.

What this family is best for: calling out lack of theory, brittleness, and interpretability gaps.
What it risks: reinforcing fatalism and giving cover for non-accountability.

3.5 Monster quotes: demons, existential dread, runaway agency

Monster quotes treat AI as a potential agent of catastrophe. The language escalates: demons, summoning, extinction, runaway intelligence, doomsday. The weirdness is deliberate. It aims to generate fear strong enough to overcome complacency, incentives, and institutional inertia.

This family functions less as a description of current systems and more as a policy intervention about trajectories. Monster quotes compress an argument about scaling, autonomy, and misalignment: if systems gain capability faster than control, the result could be catastrophic. This family often frames AI as a new class of risk, comparable to nuclear weapons or engineered pathogens, and argues for precautionary governance.

Monster metaphors can be valuable when they spur serious risk management: audits, containment practices, staged deployment, red-team culture, security hardening, and explicit accountability for misuse. Fear can be an adaptive emotion when the baseline is naive trust.

But monster quotes also degrade discourse when they collapse nuance into apocalypse. They can become self-paralyzing. They can be used to justify censorship theater, performative regulation, or opportunistic power grabs under the banner of safety. They can also cause the public to tune out: constant existential dread becomes background noise. In some contexts, monster quotes act as hype: they imply the system is almost godlike, which can draw investment and prestige.

What this family is best for: forcing attention to catastrophic risk and long-horizon governance. What it risks: polarization, fear fatigue, and hype-through-terror.

3.6 Tool quotes: “augmentation,” “adjunct,” “calculator,” “prosthesis”

Tool quotes insist that AI is best understood as an instrument that extends human capability. Instead of “AI as agent,” this family frames “AI as augmentation,” “AI as adjunct,” “AI as calculator,” “AI as prosthesis,” “AI as copilot.” The weirdness here is softer: it rejects both mystique and monster narratives and replaces them with ergonomics and workflow.

Tool quotes often show up among practitioners who care about reliable use: they want to reduce anthropomorphism and focus on how humans can supervise, verify, and integrate. The tool frame also has ethical traction: tools have operators; operators have responsibilities. Tool metaphors naturally lead to questions like: Who is accountable for outcomes? What training is required? What audits are necessary? What guardrails are baked into the interface?

The weakness of tool quotes is that they can underplay the agency illusion produced by conversational interfaces. Tools that talk back can feel like agents. Tool quotes can also understate systemic risks: when a tool is deployed at scale inside institutions, it becomes part of

governance, not merely a personal aid. A prosthesis used by millions can reshape labor markets, education norms, and information ecosystems.

What this family is best for: responsible deployment, accountability, and practical integration.
What it risks: minimizing systemic externalities and overtrusting the human-in-the-loop as a guarantee.

3.7 Deflation quotes: "just statistics," "just autocomplete," "just curve fitting"

Deflation quotes are the most common in everyday debate because they feel like clarity. They reduce AI to something unromantic: statistics, curve fitting, autocomplete, optimization. Their rhetorical mission is to puncture hype by insisting that nothing fundamentally new is happening, only scale and engineering.

This family has genuine explanatory value. It reminds audiences that models are trained from data, that outputs are probabilistic, that competence is not evidence of consciousness, and that the system's "knowledge" is not the same as a human's lived understanding. Deflation quotes can also prevent dangerous category errors: people delegating moral judgment to a system that merely predicts plausible text.

But the word "just" is doing the ideological work. "Just statistics" can be true in one sense and still miss the reality that statistical learning at scale can reorganize whole industries. "Just curve fitting" can be technically narrow while ignoring the ways learned representations support broad transfer. Deflation quotes often fail as forecasts: they underestimate the consequences of "mere" scaling and "mere" optimization.

This family also tends to produce unproductive stalemates. The pro side says, "Look what it can do." The deflation side says, "But it's just X." Both can be simultaneously correct because the disagreement is often about what counts as intelligence, what counts as understanding, and what kinds of risk matter.

What this family is best for: resisting anthropomorphic overclaiming and reminding readers about probabilistic machinery.
What it risks: complacency, underestimation, and a false sense that explanation equals dismissal.

3.8 Inversion quotes: "the stupidity of AI" and anti-genius framings

Inversion quotes reverse the halo. They treat AI not as superhuman intellect but as a system that is bizarrely, dangerously dumb in ways that matter. "The stupidity of AI" is a flagship inversion: it suggests that AI failures are not small glitches but structural mismatches between pattern competence and real-world requirements.

The inversion family is often rooted in deployment reality. Practitioners see systems that ace benchmarks yet fail on simple edge cases, that hallucinate citations, that are brittle under distribution shift, that can be manipulated by prompt injection, that behave inconsistently across near-identical inputs. Inversion quotes highlight an important truth: competence is not uniform, and failure can be sharp, surprising, and uncalibrated.

This family is also a critique of a specific kind of prestige fraud: the tendency to interpret performance on curated tasks as evidence of general intelligence. Inversion quotes insist on the non-negotiable: the world is messy, adversarial, and full of incentives to exploit system weaknesses.

The risk is that inversion can become contempt, and contempt can become blindness. Dismissing systems as "stupid" can ignore real capability gains and lead to underpreparedness. An institution that assumes "it's dumb" may fail to defend against misuse, manipulation, or subtle harms, because it mistakes brittleness for harmlessness.

What this family is best for: highlighting brittleness, edge cases, security risks, and the gap between benchmarks and deployment.
What it risks: missing real capability and creating a false sense of safety.

Across all eight families, the same meta-lesson holds: weird quotes are lenses. Each lens sharpens one dimension and distorts another. The goal is not to pick a favorite quote-family, but to learn which lens to use for which question, and to refuse the lazy move of treating any single lens as the whole truth.

4. Case Studies: What the Famous Weird Quotes Get Right and Wrong

The point of case studies is not to "score" quotes as correct or incorrect. A weird quote is a lens: it sharpens one dimension and blurs another. What matters is knowing which question the quote helps answer, and which questions it quietly breaks.

4.1 "Artificial cleverness": misnomer as a theory of understanding

This quote-family draws a bright line between competence and understanding. It says: whatever these systems are doing, do not confuse it with what humans mean by comprehension, insight, or grasp. What it gets right is the warning against category error. Many AI systems can produce plausible language, plans, or explanations without any stable internal commitment to truth, reference, or lived meaning. They can pass through the surface of language without touching the anchor points that humans use to connect words to the world.

The phrase also exposes a social hazard: the label “intelligence” invites deference. Once a tool is granted the aura of mind, people begin to treat outputs as judgments rather than artifacts. That shift is not merely philosophical. It changes who gets blamed when something fails. “The AI decided” is the modern equivalent of shrugging at a black box.

Where the “cleverness” framing can mislead is in what it quietly implies about capability. Cleverness sounds like party tricks: narrow, brittle, superficial. But large-scale systems can be clever in ways that are operationally consequential: they can draft code, summarize complex texts, search solution spaces, tutor, and coordinate actions across tools. Even if you think they lack understanding in a deep sense, they can still function as powerful cognitive instruments. A misnomer quote clarifies the metaphysics but can obscure the practical stakes.

4.2 “Neither artificial nor intelligent”: materials, labor, and ideology

This quote attacks two myths at once. “Not artificial” points to the physical substrate: the minerals, energy, servers, supply chains, water, land, and logistics that make computation real. “Not intelligent” points to the social substrate: the humans who generate the training data, label datasets, write code, curate interfaces, and clean up outputs through evaluation and feedback. The phrase is not just deflationary; it is materialist. It drags AI back down into the world of extraction and labor.

What it gets right is that “AI” is a conveniently clean abstraction. It hides both upstream costs and downstream accountability. It also hides how much of modern AI performance depends on human scaffolding: data collection choices, labeling decisions, prompt patterns, safety policies, and product constraints. The quote is a corrective to the tendency to treat AI as an autonomous force rather than an engineered and financed project.

Where it can mislead is by compressing many different systems into a single indictment. Some AI systems are closer to “automation” than “intelligence,” but others exhibit nontrivial generalization and planning-like behavior under scaffolding. Also, the quote can be interpreted as saying the technology is nothing but exploitation and ideology, which can crowd out technical evaluation. A reader can hold both truths at once: AI is materially intensive and socially organized, and it still creates real new capabilities that require serious governance.

4.3 “Stochastic parrots”: fluency, grounding, and social risk

The “parrot” metaphor targets a specific failure mode: systems that produce fluent language without grounded meaning. It warns that language competence is not the same as understanding, and that high-quality imitation can be socially dangerous because humans trust coherence. The metaphor also points to a structural risk: models trained on massive text corpora can absorb biases, reproduce stereotypes, and amplify harmful narratives, all while sounding confident and polite.

What it gets right is the emphasis on social impact and epistemic hygiene. Fluency can bypass skepticism. It can create a false sense of authority, especially when deployed in education, medicine, law, and bureaucracy. The metaphor also reminds us that output plausibility is not truth. A system can "sound like" it knows without knowing.

Where it can mislead is by implying that the system is merely replaying memorized phrases. Modern language models do sometimes regurgitate, but much of their usefulness comes from abstraction and recombination. The parrot image is strongest as a critique of grounding and intent, not as a denial of all novelty. Another limitation is that the metaphor can become a moral label rather than an analytic tool. Once something is called a parrot, people stop asking which tasks are safe, which tasks are unsafe, and what guardrails make sense.

4.4 "Blurry JPEG of the web": compression, generalization, and failure modes

This metaphor frames language models as lossy compression. They learn to store the gist of internet-scale text and then "decompress" it into plausible outputs. The JPEG analogy predicts exactly the kinds of artifacts people observe: smoothness where sharp edges should be, plausible details that are wrong, and distortions that become obvious under zoom. It is a powerful explanation for hallucinations: when the model lacks a high-resolution memory of a specific fact, it fills in with statistically plausible texture.

What it gets right is the distinction between style and truth. Compression preserves patterns but discards details, and the discarded details matter most when you ask for citations, exact numbers, or niche facts. It also clarifies why these systems can be impressive at summarization while unreliable at precision. The metaphor helps readers stop treating outputs as retrieved records and start treating them as synthesized reconstructions.

Where it can mislead is by suggesting the model is only a degraded copy of the dataset. Compression can enable generalization. A good compressor captures structure, not just pixels. In language, structure includes grammar, argument patterns, conceptual relationships, and procedural knowledge embedded in text. The JPEG metaphor explains error modes well, but it can understate the extent to which learned structure can support new problem-solving behavior, especially when combined with tools and verification steps.

4.5 "Alchemy": engineering without theory and the limits of interpretability

"Alchemy" is a critique from within technical culture. It says: the field can make things that work, but often cannot explain why they work, cannot reliably predict failure, and cannot provide guarantees commensurate with high-stakes use. In other words, it accuses modern ML practice of being recipe-driven: hyperparameters, tricks, scaling, and empirical wins without deep theory.

What it gets right is the gap between performance and understanding. Many real failures in deployed ML systems come from distribution shift, adversarial manipulation, hidden data leakage, and brittle generalization. “Alchemy” captures the uncomfortable truth that success can outpace comprehension, and that opacity can be socially unacceptable when systems influence livelihoods, health, liberty, or war.

Where “alchemy” can mislead is by implying the field is irrational or pre-scientific. Engineering often advances before theory. Many domains matured through a phase where empirical practice led the way and explanation followed. Also, the critique can be too broad: parts of ML are deeply theory-informed, and interpretability, evaluation science, and robustness research have advanced significantly. “Alchemy” is most useful as a warning against complacency and overclaiming, not as a blanket dismissal of the discipline.

4.6 “Curve fitting”: representation, causality, and where it becomes misleading

“Curve fitting” is a deflationary phrase that often means: these models learn correlations, not causes. They interpolate within the space defined by training data; they do not truly understand mechanisms. The phrase targets a real limitation: purely observational training struggles with causal inference, counterfactual reasoning, and robust transfer to new regimes. It also highlights why models can fail confidently when the environment changes: they learned the curve, not the engine.

What it gets right is that correlation-based learning has sharp limits in domains where causes matter more than patterns. In medicine, policy, and physical systems, causal structure determines what happens when you intervene. A model that merely mimics historical patterns can encode the past, not the truth.

Where “curve fitting” becomes misleading is when it equates all representation learning with trivial interpolation. High-dimensional curve fitting can discover abstractions that behave like concepts. It can support nontrivial generalization, reasoning-like behavior, and tool-augmented workflows. Also, “curve fitting” can be weaponized as a rhetorical off-switch: it invites the listener to stop thinking. A better use is conditional: ask which tasks require causal guarantees and which tasks tolerate pattern-based assistance.

4.7 “Autocomplete on steroids”: UI metaphors and underclaim/overclaim traps

This metaphor anchors the system in next-token prediction and reminds readers that, at the core, the model generates continuations. It is effective because it demystifies. It warns that the model is not consulting a database of truth; it is producing plausible text. It also predicts a familiar failure: when the prompt implies a false premise, the system may continue it smoothly rather than challenge it. That is autocomplete behavior, scaled up.

What it gets right is the interface-level truth: conversational AI often behaves like a completion engine. Users can be seduced by fluency into believing they are interacting with a mind, when they are interacting with a generator conditioned on context. The metaphor also surfaces a critical governance point: if the UI looks like a trusted advisor, users will treat it like one. The boundary between suggestion and decision blurs.

Where it can mislead is by implying the system cannot do more than completion. With tool use, external memory, verification loops, and structured prompting, these systems can perform multi-step tasks that feel qualitatively different from autocomplete. The phrase can become an underclaim that blinds institutions to real capability and real risk. It also creates an overclaim in the opposite direction: by calling it "on steroids," it implies an exponential leap without specifying where the leap actually comes from (scale, data, training regime, reinforcement, scaffolding, tools). The best reading is: the core mechanism is continuation, but the emergent behavior can exceed what the metaphor invites you to expect.

4.8 "AI effect": why working systems stop being called AI

The "AI effect" is the observation that once a technique becomes reliable and routine, people stop calling it AI and start calling it software. Chess-playing computers were once AI; now they are programs. Speech recognition, machine translation, recommender systems, fraud detection, route planning: each was once "AI" when it felt magical, and later became infrastructure when it became mundane.

What this framing gets right is that "AI" is partly a moving label for the frontier of what seems mind-like. The label tracks social surprise more than technical essence. This explains why public debates about AI can feel incoherent: people are not arguing about a stable category. They are arguing about what still feels uncanny, and about which tasks we are willing to treat as "understood."

Where the AI effect can mislead is by encouraging a cynical stance: that "AI" is only marketing. Marketing is part of it, but the effect also reflects a real cognitive phenomenon: humans quickly normalize capabilities once they become tools. That normalization can be dangerous if it leads to under-regulation or under-scrutiny. A system can become boring and still be powerful enough to reshape labor, education, and politics. The AI effect is therefore not only about semantics. It is a warning about habituation: society can stop noticing the transformative thing right as it becomes ubiquitous.

Taken together, these case studies show why weird quotes persist. They are not mere entertainment. They are compressed arguments about understanding, materials, labor, risk, and responsibility. The reader's job is not to pick one quote as the final truth, but to recognize which

quote-family fits which question, and to refuse the bait of slogans that try to do all the thinking for you.

5. Why Smart People Speak This Way

At first glance, weird quotes about AI can look like intellectual theater: clever people competing for the sharpest line. But the pattern is deeper. Smart people reach for compressed metaphors and provocative inversions because the surrounding environment makes full explanations unusually hard. AI sits at the intersection of technical complexity, public fear, commercial incentives, and unresolved philosophical questions. In that terrain, careful prose often loses to portable phrasing. A memorable one-liner becomes the carrier wave for an argument that otherwise would never reach the audience.

The oddity is not that smart people talk this way. The oddity is that they are expected not to. When the public conversation is fast, adversarial, and shaped by incentives that punish nuance, rhetoric becomes a survival tool. Sometimes it clarifies. Sometimes it degrades. Often it does both at once.

5.1 Communication constraints: interviews, headlines, and virality

Most weird AI quotes are not produced in ideal scholarly conditions. They are produced under constraints that virtually guarantee compression.

Interviews are time-boxed. A researcher or critic is asked to explain a sprawling field in seconds. The interviewer wants a crisp takeaway, not an epistemology seminar. The speaker knows that the audience's attention is scarce and that the conversation will likely be edited. Under these constraints, the speaker chooses a line that can stand alone if the surrounding context gets cut. That line is engineered for survivability.

Headlines reward polarity. Editors are selecting for what will be clicked, remembered, and shared. A sentence that contains caveats and conditional statements does not travel. A sentence that contains an image does. A sentence that contains a conflict does. "Blurry JPEG" travels. "Stochastic parrot" travels. "Summoning the demon" travels. The selection pressure is obvious: quotes evolve to fit the ecosystem that propagates them.

Virality favors metaphor over mechanism. Humans share images, not diagrams. If you explain the technical reason a model hallucinates—distributional uncertainty, lack of retrieval grounding, reinforcement shaping, and so on—you will lose most listeners. If you offer "lossy compression," you give them a mental picture that fits in working memory. Even if it is only partially accurate, it is cognitively affordable.

There is also the problem of mixed audiences. An AI expert may be speaking simultaneously to non-experts, policymakers, investors, opponents, and fans. Each group brings different priors and different anxieties. The speaker picks a phrase that can be decoded by multiple audiences at once, even if each audience decodes it differently. This is one reason weird quotes become ambiguous: ambiguity is sometimes the only way to communicate across audience heterogeneity.

Finally, there is a harsh asymmetry: a careful explanation can be refuted by a single vivid counterexample, while a vivid metaphor can survive many counterexamples because it is interpreted elastically. If a model solves a hard task, "it's just autocomplete" can still be preserved by shifting the definition of "just." If a model fails a simple task, "the stupidity of AI" can still be preserved by pointing to brittleness. Weird quotes are robust to empirical whiplash. That robustness makes them attractive under fast-moving conditions.

5.2 Incentives: funding, warning, branding, and reputational defense

Communication constraints explain how weird quotes spread. Incentives explain why they are produced in the first place.

Funding incentives reward differentiation. Researchers, labs, and institutions compete for attention, grants, and prestige. A distinct framing can function like a research brand. If you can coin a metaphor that captures a widespread unease or insight, you gain narrative leverage. The phrase becomes a shorthand for your position, and people cite the phrase as they cite you. In a crowded discourse space, a strong line is a competitive advantage.

Warning incentives reward alarm. If you sincerely believe a technology may cause serious harm, you face a moral dilemma: speak in careful measured language and risk being ignored, or speak in a way that shocks attention into action. Many monster metaphors come from this situation. They are not necessarily theatrical for its own sake; they are an attempt to solve the "attention deficit" of precaution. Whether the strategy is effective is another question, but the incentive to escalate is obvious when baseline attention is low.

Branding incentives reward enchantment. Commercial actors benefit when AI feels magical, inevitable, and world-historic. Even when companies use deflationary language in technical documentation, their marketing often leans toward anthropomorphic framing because it sells. In response, critics use weird quotes as counter-branding: "not intelligent," "just statistics," "parrots," "blurry JPEG." The weirdness is a counterspell against enchantment.

Reputational defense incentives reward distance. When a system fails publicly, many actors want to minimize blame: "the model did something unexpected," "it's a black box," "it hallucinated," "emergent behavior." These phrases can be descriptive, but they can also function as rhetorical shields. Weird quotes that emphasize opacity can serve as preemptive liability

management: they remind audiences that surprises are possible, which can reduce outrage when surprises occur.

There is also intradisciplinary politics. Different research traditions—symbolic AI, statistical learning, causal inference, embodied cognition—compete for influence. A deflationary quote from one camp can be a strategic critique of the other camp's claims. "Just curve fitting" is not only about the public; it is also about arguing that one paradigm is incomplete. "Alchemy" can be a critique of a culture that prioritizes benchmarks over understanding. These quotes are weapons in internal debates, not just public education.

Incentives do not make quotes dishonest. They make quotes selected. They shape which truths get amplified and which truths get softened.

5.3 Philosophical priors: mind, meaning, consciousness, and computation

Some weird quotes are downstream of deep philosophical commitments. People do not agree on what "intelligence" is, what "understanding" is, what "meaning" is, or whether consciousness is relevant to moral status and epistemic authority. When those foundations differ, the same observed behavior yields different interpretations, and those interpretations crystallize into quotable phrases.

If you believe intelligence requires grounded semantics—reference to the world, embodiment, intention—then language models look like fluent simulators. A parrot metaphor feels natural. If you believe intelligence is primarily functional—competent behavior across tasks—then the same systems look like early forms of general intelligence, and deflation metaphors feel like denial.

If you believe consciousness is non-computational or tied to particular physical processes, then you will resist "intelligence" language for purely digital systems. If you believe consciousness is emergent from computation, you may be more open to mind-like interpretations. These priors are rarely stated in popular debate, but they silently drive which weird quote feels "obviously true."

Meaning itself becomes a battlefield. Some views treat meaning as use: if a system uses language effectively, it participates in meaning. Others treat meaning as anchored in experience: without lived context, there is no real meaning. This is why "understanding" becomes the pivot word of many quotes. People talk past each other because they use "understanding" to mean different things, then treat the disagreement as empirical rather than definitional.

Weird quotes function as philosophical flags. They let speakers plant a position without writing a philosophy paper. The cost is ambiguity. The benefit is speed.

5.4 The pedagogical motive: metaphors as scaffolding for non-experts

Not all weird quotes are weapons. Many are teaching tools.

AI is difficult to explain because its core mechanisms are both conceptually simple and behaviorally complex. “Next-token prediction” is easy to say, but it does not predict what happens when you scale that prediction across huge datasets and add reinforcement, tool use, and interaction loops. Non-experts need scaffolding: intermediate metaphors that let them hold a partial model while they learn.

Metaphors serve three pedagogical functions. First, they reduce cognitive load. “Lossy compression” is easier to carry than an explanation of probabilistic inference and uncertainty calibration. Second, they cue appropriate skepticism. “Blurry JPEG” warns you not to treat fine details as reliable. “Parrot” warns you not to mistake fluency for wisdom. Third, they guide safe use. “Tool not creature” encourages verification and accountability. “Calculator” encourages checking. “Prosthesis” encourages thinking about human dependence and skill atrophy.

The danger is that scaffolding can become a permanent building. People cling to the metaphor long after it has served its purpose, and the metaphor becomes ideology. “Just autocomplete” becomes a refusal to see new capabilities. “Demon” becomes a refusal to see mundane governance solutions. The pedagogical motive can backfire when the metaphor is remembered but the lesson is not.

The best pedagogical weird quotes are those that are explicitly framed as partial: “This is a way to think about one aspect.” They invite the reader to ask what the metaphor does not cover. In practice, however, the social media environment strips away that humility. The metaphor becomes an identity marker: which team you are on.

Ultimately, smart people speak this way because weird quotes are adaptive under modern conditions. They compress, they travel, they frame, they warn, they teach, and they compete. The task for the reader is to recover what has been compressed: the assumptions, the incentives, the intended audience, and the specific failure modes that the quote was built to highlight.

6. What Metaphors Illuminate and What They Hide

Metaphors are not decorations; they are cognitive instruments. They are how humans compress complexity into something portable enough to reason with. That is why weird quotes about AI matter: they are the metaphors that have “won” the cultural selection process. They circulate because they help people make sense of systems that are otherwise too technical, too fast-moving, and too socially consequential to hold in mind.

But every metaphor is a bargain. It grants clarity on one axis and charges you in blindness on another. A metaphor illuminates by projecting structure from a familiar domain onto an unfamiliar one. The projection is never perfectly faithful. It smuggles in assumptions. It pushes attention toward certain questions and away from others. Over time, the metaphor becomes not merely a way of speaking, but a way of seeing, and then a way of deciding.

This section maps that bargain explicitly. For each metaphor family we ask: what does it highlight, what does it hide, and how does it shape the agency we attribute to AI systems and the responsibility we assign to humans.

6.1 What each family highlights: capability, risk, or epistemic limits

Different metaphor families are optimized for different kinds of illumination. Some are about capability. Some are about risk. Some are about epistemic limits, meaning: how we should interpret the output and what kind of knowledge claims we are allowed to infer from it.

Misnomer metaphors highlight epistemic boundaries. “Not intelligence, just cleverness” shines a light on the difference between producing competent outputs and possessing understanding, intention, or wisdom. It attempts to block a specific error: granting authority to a system because it sounds mind-like. This lens is particularly useful when a system is being positioned as an advisor, judge, or moral voice. It insists: do not confuse the appearance of reason with the presence of reasons.

Compression metaphors highlight a specific failure mode: loss of detail. “Blurry JPEG” tells you to trust coarse structure more than fine grain. It helps readers predict hallucinations as artifacts of lossy reconstruction rather than as malicious lies or mystical spontaneity. Compression metaphors also illuminate why models are good at paraphrase, summarization, and stylistic transformation, but unstable at precise citation, exact numbers, and niche factual claims.

Mimicry metaphors highlight social risk and meaning risk. “Parrot” makes vivid the possibility of language without grounding. It also emphasizes the danger of persuasion-by-fluency: humans tend to trust coherent narrative. Mimicry metaphors are good at reframing the system as a simulator rather than an authority. They are especially potent for warning against uncritical use in education, therapy, law, and politics, where trust and meaning are central.

Magic metaphors highlight epistemic fragility and governance difficulty. “Alchemy” and “black box” emphasize that performance can be real even when explanations are weak and guarantees are scarce. They are a reminder that success in benchmarks does not necessarily equal predictable behavior under distribution shift, adversarial pressure, or institutional incentives. Magic metaphors illuminate the sense in which a system can be practically useful while scientifically under-explained, which is exactly where governance gets hard.

Monster metaphors highlight existential risk and urgency. They spotlight tail risks, runaway dynamics, and the fear that capability scaling can outrun control. Whatever one thinks of the probability, the metaphor's function is to force attention onto worst-case governance problems that are otherwise easy to postpone. Monster metaphors are often not about current systems; they are about trajectories, incentives, and irreversible harm.

Tool metaphors highlight practical integration and accountability. "Copilot," "calculator," "prosthesis," "adjunct" are metaphors for how people should use AI: as an extension under supervision. These metaphors illuminate workflows, verification, and the division of labor between human and machine. They are strongest in domains where bounded assistance is valuable and where verification is feasible.

Deflation metaphors highlight mechanism-level humility. "Just statistics," "just curve fitting," "just autocomplete" push attention back to the training process and away from metaphysical speculation. They are useful for puncturing a specific kind of mystique: that the system is a new mind or a moral actor. Deflation metaphors emphasize that the system is built from data, optimization, and architecture choices, not divine inspiration.

Inversion metaphors highlight brittleness and adversarial vulnerability. "The stupidity of AI" draws attention to the ways systems can fail in non-human ways: confident nonsense, weird edge cases, susceptibility to prompt injection, and inconsistency. Inversions are valuable because they counterbalance the halo effect created by impressive demos. They help institutions resist the temptation to equate benchmark wins with real-world reliability.

Seen this way, each metaphor family is like a flashlight aimed at a particular feature of the landscape. The trouble begins when the flashlight is mistaken for the whole sun.

6.2 What each family obscures: data pipelines, incentives, and human labor

The consistent blind spot across most metaphor families is the socio-technical stack. Many weird quotes talk as if AI were a single object, when in reality it is a pipeline embedded in organizations.

Data pipelines are often invisible in metaphors. A model does not simply "know" things; it is trained on particular corpora, filtered by particular policies, shaped by particular reinforcement signals, and deployed through particular retrieval and tooling layers. A compression metaphor may make you think in terms of dataset-to-output mapping, but it can obscure data governance: what was included, excluded, scraped, licensed, labeled, cleaned, and curated. It can also hide the nontrivial role of post-training: alignment tuning, red-teaming, and safety constraints that shape behavior in ways that are not "in the data" in any simple sense.

Incentives are also frequently obscured. A tool deployed inside a customer-support organization behaves differently than the same model used by a hobbyist at home, because the surrounding incentives differ: metrics, time pressure, liability, and oversight. Monster metaphors can obscure corporate incentives that drive premature deployment. Deflation metaphors can obscure investor incentives that inflate claims. Magic metaphors can obscure the incentive to maintain opacity as a competitive moat. When metaphors omit incentives, they encourage the fallacy that outcomes are caused by the model alone rather than by the system-of-use.

Human labor is one of the most systematically hidden dimensions. Many metaphors treat AI as an autonomous producer, which erases the humans who wrote the data, labeled the training sets, moderated content, built guardrails, evaluated outputs, and cleaned up the failures. Even when people acknowledge that data comes from humans, they often miss the ongoing labor of making systems safe and usable: prompt engineering, monitoring, incident response, and policy enforcement. Tool metaphors can obscure this by portraying the system as a simple instrument, when in reality the instrument is continuously maintained by human operators and institutional processes.

Metaphors also obscure distribution. When we talk about "AI," we often ignore the gap between the frontier model in a lab and the model actually deployed across devices, regions, and contexts. Safety constraints, retrieval layers, and policy filters differ. The same metaphor applied globally masks local reality.

Finally, metaphors obscure the political economy of AI: compute concentration, supply chain dependencies, energy use, and geopolitical constraints. A misnomer quote about "intelligence" says nothing about the world of chips, water, grid constraints, and procurement. Yet those material constraints can determine which systems exist, who controls them, and which societies become dependent.

The rule is simple: metaphors are about meaning, but AI outcomes are about systems. Meaning without systems becomes myth.

6.3 Anthropomorphism and the agency illusion

The most consequential hidden effect of metaphor is that it shapes agency attribution.

Humans are tuned to treat language as a signature of mind. When something speaks fluently, we automatically infer a speaker behind the speech. Conversational AI exploits this ancient circuit. Even if you intellectually know the system is statistical, your social cognition keeps trying to locate an agent. Metaphors can amplify or dampen this tendency.

Monster metaphors amplify agency: demons, runaway minds, autonomous forces. Magic metaphors can also amplify agency indirectly by framing the system as mysterious and

uncontrollable. Even some positive metaphors—“copilot,” “assistant,” “teammate”—invite a subtle reclassification of the system from tool to partner. Once that happens, people begin to share responsibility with it. They accept its suggestions with less scrutiny. They say “the AI decided.” They begin to blame it or credit it.

Deflation and tool metaphors aim to dampen anthropomorphism, but they can fail because the interface remains social. The agency illusion does not require belief in consciousness. It only requires the experience of being responded to. That is enough to create a sense of dialogue, and dialogue invites moral framing: helpful, deceitful, lazy, sincere. This is why users talk about models “lying” when they hallucinate, even though lying requires intent. The language of agency rushes in because it is the natural vocabulary of conversation.

The agency illusion has direct governance consequences. When an output causes harm, who is responsible? If the system is framed as an agent, responsibility diffuses. If it is framed as a tool, responsibility concentrates on designers, deployers, and operators. This is why misnomer quotes are not merely pedantic: they are attempts to prevent a moral and legal slippage where accountability evaporates into the machine.

There is also a subtler effect: anthropomorphism changes what people ask. If you think you are talking to a mind, you ask for advice, judgment, and permission. If you think you are using a tool, you ask for drafts, options, computations, and checklists. The same system can yield very different outcomes depending on which posture the user adopts.

A mature stance is to treat agency as layered. The model does not have agency in the human sense, but the system-of-use produces agency-like effects. A deployed model can initiate actions through tools, trigger workflows, and influence decisions. It can function as an agent inside a socio-technical loop even if it is not an agent in metaphysics. That distinction is crucial: we can govern the loop without pretending the model is a person.

The central lesson of this section is that weird quotes are not trivial. They are the handles by which society grips a moving technology. If the handle is shaped wrong, people will mis-lift the object. They will either drop it in fear, swing it recklessly, or hand it authority it does not deserve. The reader’s task is to use metaphors consciously: as partial lenses, not total explanations, and always with the system-of-use in view.

7. How Weird Quotes Shape Public Policy and Institutional Behavior

Weird quotes about AI do not remain in the realm of rhetoric. They leak into institutions as operational premises. A single metaphor can become a shared shorthand inside a regulatory agency, a boardroom, a newsroom, a school district, or a military procurement meeting. Once

that happens, the quote stops being commentary and becomes a causal factor. It alters what people perceive as the object being governed, what harms they imagine, what safeguards they demand, and what responsibilities they assign. Because AI policy is still definitional in many places, the metaphors that dominate the debate often determine the shape of the rules that follow.

The deeper reason is simple: institutions are forced to act under uncertainty. They cannot wait for perfect technical clarity or philosophical consensus. They need usable concepts now. Weird quotes supply those concepts in a portable form. The danger is that portability can substitute for precision, and that a metaphor can silently become law.

7.1 Regulation: definitions that decide what gets governed

Regulation begins with a definition. You cannot govern "AI" until you decide what counts as AI. And in practice, definitional boundaries determine what is regulated, who is liable, and what compliance requires. Weird quotes influence those boundaries by either expanding the category ("AI is a demon; treat it as an existential hazard") or shrinking it ("AI is just statistics; treat it as ordinary software").

Misnomer quotes play directly into regulatory framing. If a policymaker absorbs "AI is not intelligence but cleverness," they may lean toward laws that treat AI as a form of automation: regulate deployment contexts, documentation, and accountability, but avoid metaphysical claims about agency. This can produce a governance posture focused on human responsibility and auditability. It can also lead to under-regulation if the policymaker interprets "cleverness" as triviality.

Monster quotes push the opposite way: they argue the category is so dangerous that it deserves a special regime, perhaps analogous to nuclear or biosecurity governance. This can motivate serious safety architecture—model evaluations, red-teaming, incident reporting, restrictions on certain deployment classes, and controls over compute or model weights. It can also motivate sweeping, vague restrictions that are difficult to enforce and that may inadvertently consolidate power in the hands of large incumbents.

Deflation quotes often show up when regulators are trying to avoid moral panic or overreach. "It's just curve fitting" can lead to treating AI as merely an extension of existing statistical software. That can be reasonable in some contexts, but it can also leave gaps where new capabilities change the risk profile: mass-scale persuasion, automated exploitation, and capability amplification that outstrips legacy compliance frameworks.

Tool metaphors can produce targeted governance: rules focused on use-cases rather than on "the model" as a thing. This is often more workable. The risk is that tool metaphors can mask

systemic harms. A tool used at scale can become a social infrastructure, and infrastructures require different governance than personal tools.

In other words, weird quotes are not merely opinions. They are definitional levers. When one metaphor wins, a particular style of regulation becomes thinkable, and other styles become difficult to justify.

7.2 Procurement and safety: when slogans replace evaluation

Procurement decisions—whether in government, education, healthcare, or enterprise—are where metaphors become budgets, and budgets become reality. Procurement is often made under time pressure, marketing pressure, and executive pressure. In that environment, slogans can become substitutes for evaluation.

If decision-makers adopt the "autocomplete" metaphor, they may treat AI as a text-generating convenience tool and underestimate the need for governance. They may deploy it widely without robust verification workflows, human oversight, or security controls. They may assume errors will be obvious and benign. In reality, some errors are subtle, plausible, and costly. An "autocomplete" framing can lead to insufficient monitoring and inadequate incident response.

If decision-makers adopt the "black box" or "alchemy" metaphor, they may overestimate unpredictability and either refuse adoption entirely or treat adoption as inevitable risk with no middle path. Both outcomes can be harmful. Refusal can deprive institutions of tools that could improve service and reduce burdens. Fatalistic adoption can produce a governance vacuum: if it is a black box, why bother evaluating? The black box metaphor can become an excuse for non-accountability.

If decision-makers adopt "AI is neither artificial nor intelligent," they may be pushed to consider supply chain ethics, data provenance, energy costs, labor impacts, and externalities. This is a healthy corrective to narrow capability-focused procurement. The risk is that the institution may treat the quote as a moral verdict rather than a governance checklist.

Procurement is where "evaluation discipline" matters most, because this is where systems meet adversarial reality. Yet evaluation is hard and expensive. It requires test suites, red teaming, security review, privacy review, model cards, vendor transparency, and human factors analysis. Institutions are tempted to skip this and lean on metaphors instead. When slogans replace evaluation, the organization becomes vulnerable to predictable failure modes: hallucinations treated as facts, biased outputs treated as neutral, and prompt-injection attacks treated as user error.

A mature procurement posture treats weird quotes as prompts for evaluation, not as evaluation themselves. "Blurry JPEG" should trigger: what retrieval and citation mechanisms do we have?

“Stochastic parrot” should trigger: what bias and safety audits exist, and what human review is required? “Alchemy” should trigger: what interpretability and reliability evidence can the vendor provide? “Tool not creature” should trigger: what accountability chain is defined?

7.3 Moral panic vs moral anesthesia: fear that blinds, comfort that blinds

Weird quotes generate emotional posture. In policy and institutional settings, emotional posture can be decisive. The problem is not emotion; the problem is the wrong emotion dominating.

Monster quotes can generate moral panic. Panic is not the same as prudence. Panic narrows attention to catastrophic narratives and can lead to policy that is theatrical rather than effective. It can justify blanket bans, poorly defined restrictions, or politically opportunistic control measures that have little relationship to real risk. It can also create fear fatigue: if every new model is framed as an apocalypse, audiences eventually stop listening.

Deflation and tool metaphors can generate moral anesthesia. Anesthesia is not the same as calm. Anesthesia numbs institutions to slow-moving harms: labor displacement, concentration of power, surveillance, manipulation, and the quiet erosion of expertise. If AI is “just software,” the institution may treat the technology as unremarkable and fail to build oversight commensurate with its social reach. It may normalize the delegation of sensitive decisions to systems that do not deserve that role.

The deeper risk is that institutions oscillate between panic and anesthesia, never settling into disciplined governance. A scandal triggers panic; then vendors promise safety; then adoption accelerates under anesthesia; then another scandal triggers panic again. This oscillation is not a failure of intelligence; it is a failure of stable frameworks and incentives.

Weird quotes can either fuel the oscillation or help stabilize it. Used wisely, they can map the extremes: panic metaphors warn of tail risks; deflation metaphors warn of category errors. The institutional goal is to hold both: to be neither mesmerized nor dismissive.

7.4 Media dynamics: headline selection and narrative path dependence

Media is the habitat in which weird quotes evolve. The media logic selects for lines that create story tension: genius versus fraud, salvation versus doom, magic versus stupidity. Once a metaphor becomes a recurring headline trope, it shapes narrative path dependence: future stories are framed through the same lens because editors and audiences already recognize it.

Path dependence matters because it affects what questions journalists ask and what answers sources supply. If the dominant narrative is “AI as monster,” then interviews focus on existential risk, regulation, and dramatic predictions. If the dominant narrative is “AI as autocomplete,” then interviews focus on product features, productivity, and incremental improvement. If the

dominant narrative is “AI as plagiarism,” then interviews focus on copyright and training data. Each narrative selects different experts and different evidence. Over time, the public thinks it is seeing the whole landscape, but it is seeing a curated corridor.

Headline selection also encourages quote laundering. A careful statement with qualifiers becomes a definitive-sounding line. The line circulates without its boundary conditions, and it becomes an ideology token. People use it to signal membership in a camp, and camps become more important than truth-seeking. The quote ceases to be a lens and becomes a flag.

The media dynamic also shapes institutional behavior indirectly. Regulators read the same headlines as everyone else. Boards and executives are sensitive to reputational risk. Universities respond to public fear. Agencies respond to political pressure. A meme-like quote can therefore function as an accelerant: it can speed up legislation, trigger procurement freezes, or provoke moral outrage that demands immediate action.

A realistic strategy is not to hope for media to become slow and nuanced. The strategy is to design institutions that can withstand narrative shocks: stable evaluation protocols, public documentation, audit structures, and independent expertise. Weird quotes will continue to surge. The question is whether governance will be driven by surges or by structures.

The practical conclusion of this section is that weird quotes are not merely speech. They are policy inputs. They shape definitions, they shape procurement, they shape emotions, and they shape the media narrative loop that feeds back into institutions. The proper response is not to ban metaphors. It is to use metaphors deliberately: to translate them into operational questions, and to ensure that the decisions they influence are anchored in evidence, incentives, and accountability chains rather than in slogans that merely feel true.

8. A Framework for Readers: How to Evaluate Weird Quotes Without Becoming Cynical

The modern information environment trains people into two bad habits. The first is credulity: treat a quotable line as wisdom because it is vivid and delivered by a prestigious person. The second is cynicism: treat all quotable lines as branding, manipulation, or ideological noise. Both habits are lazy, and both are dangerous. Credulity hands authority to slogans. Cynicism hands power to whoever acts while everyone else shrugs.

A better stance is disciplined interpretation. A weird quote is neither an argument nor a joke; it is an argument compressed into a mnemonic. Your job as a reader is decompression. You want to recover the hidden parameters: context, claim, mechanism, scope, and stakes. Once you do that, the quote stops being magic. It becomes something you can evaluate, disagree with, refine, or translate into sensible action.

The framework below is designed as a set of tests. Each test is fast enough to apply in real time, but strong enough to prevent slogan-driven thinking.

8.1 Context test: who said it, where, and in response to what

Start by reconstructing the scene. A quote is an action inside a conversation. Without the conversation, you cannot know what the quote was doing.

Ask five questions:

Who said it? A researcher speaking within their domain, a journalist writing for a general audience, an executive marketing a product, a critic advancing an ethical warning, a philosopher making a metaphysical claim. The same sentence means different things depending on speaker role and incentives.

Where was it said? A peer-reviewed paper, an academic lecture, a panel, a podcast, a tweet, a courtroom, a marketing keynote. Medium shapes message. A stage rewards drama; a paper rewards precision; a podcast rewards narrative and personality.

In response to what? Many "weird quotes" are answers to leading questions. "Is AI conscious?" "Is this dangerous?" "Is this real intelligence?" The question often contains the framing, and the quote is a counter-framing. If you don't know the question, you may misread the quote as a general statement rather than a targeted rebuttal.

When was it said? Date matters. The technology, the discourse, and the incentive landscape change quickly. A skeptical quote from an earlier era can become misleading if applied to today's systems. A hype quote from a earlier era can become embarrassing if repeated today.

To whom? Audience is decisive. A quote aimed at policymakers will emphasize accountability and harm. A quote aimed at engineers will emphasize mechanism and failure modes. A quote aimed at investors may emphasize opportunity or urgency.

Passing the context test does not prove a quote true. It prevents the first error: treating an extracted phrase as a timeless axiom.

8.2 Claim test: what is the falsifiable core (if any)

Once you have context, ask: what is the quote actually claiming? Strip away the vibe. Translate the metaphor into plain propositions.

Many quotes contain at least one falsifiable claim. "It hallucinates" can be tested: does it produce confident false statements under certain prompts? "It is lossy compression" can be tested indirectly: does it behave like a system optimizing for predictive likelihood, producing plausible reconstructions rather than retrieving exact records? "It is just curve fitting" can be

tested: does performance collapse under distribution shift, and does it fail on causal interventions?

Other quotes are not falsifiable in the strict sense because they are philosophical or normative. “It doesn’t understand” depends on what counts as understanding. “It’s not intelligent” depends on what counts as intelligence. These can still be meaningful, but they are not the same kind of claim as “it makes factual errors.” In those cases, look for the operational surrogate. If someone says “no understanding,” they may mean “no reliable truth-tracking,” or “no grounded reference,” or “no stable self-model,” or “no agency with responsibility.” Those surrogates can be tested.

A quote can also contain a non-falsifiable rhetorical move with a falsifiable tail. “This is dangerous” is not falsifiable as a general claim, but “this increases risk of fraud at scale,” “this enables prompt-injection failures,” or “this can automate targeted persuasion” are falsifiable in narrower terms.

The goal is to extract a core that can be argued about with evidence rather than with teams and vibes.

8.3 Mechanism test: what must be true technically for the quote to hold

Now ask: what model of the system does the quote assume? Every metaphor implies a mechanism. If the implied mechanism is wrong, the quote may still be emotionally appealing but technically misleading.

For each quote, identify the implied machinery:

If the quote is “autocomplete,” it assumes next-step continuation dominates behavior. Mechanism questions: is the system purely generative, or is it using retrieval, tools, and verification loops? Is it trained only on next-token prediction, or also shaped by reinforcement, preference modeling, and safety constraints? The more scaffolding exists, the less complete the “autocomplete” mechanism is as an explanation.

If the quote is “blurry JPEG,” it assumes lossy reconstruction rather than lookup. Mechanism questions: does the system have access to external memory? Is the output grounded by citations? Does it have retrieval-augmented generation? If yes, the JPEG metaphor still explains the base generator, but may not describe the full system in use.

If the quote is “curve fitting,” it assumes correlation learning without causal structure. Mechanism questions: is the model used in a predictive regime where correlation is sufficient, or in an intervention regime where causality matters? Does the task require counterfactual reasoning? Is there a causal model layer or structured reasoning component? Even if the base model is correlational, the deployment system can add causal constraints.

If the quote is “stochastic parrot,” it assumes mimicry without grounding. Mechanism questions: is the model connected to sensors, environments, or tools? Is it trained with feedback from real-world interactions? Does it have consistent world models or only statistical associations? “Parrot” is most accurate when the model’s only anchor is text.

If the quote is “alchemy” or “black box,” it assumes unpredictability and lack of interpretability. Mechanism questions: what evaluation regime exists? How well is behavior characterized under stress? Does the organization have red-team results, incident logs, and reproducibility tests? “Black box” can be a lazy label if rigorous evaluation exists, but it can also be a valid warning when evaluation is thin.

The mechanism test protects you from a common error: treating a metaphor about the core model as if it describes the whole deployed system, or treating a metaphor about one failure mode as if it describes all behavior.

8.4 Scope test: which tasks, settings, and time horizons are implied

Most quote fights are actually scope fights. People argue as if the quote is universal when it was meant to be conditional.

Ask: scope over what?

Task scope: Is the quote about creative writing, coding, medical advice, legal interpretation, strategic planning, persuasion, or scientific discovery? A model can be “blurry JPEG” for factual recall and “surprisingly useful” for brainstorming. It can be “autocomplete” in prose and “capable” when paired with tools.

Setting scope: Is it an individual user, a classroom, a hospital, a call center, a military system, a social network? The same model can be relatively safe in one setting and catastrophic in another because of scale, incentives, and consequences.

Adversarial scope: Is the environment cooperative or adversarial? Many systems look fine until someone tries to exploit them. Prompt injection, jailbreaks, data poisoning, and targeted manipulation change the landscape.

Time horizon scope: Is the quote about current systems or future trajectories? Monster metaphors often jump time horizons. Deflation metaphors often freeze the present and deny plausible acceleration. Ask explicitly: is this claim about what exists now, or what may exist after several capability jumps?

Population scope: Who is affected? A tool may help a skilled expert and harm a novice who overtrusts it. It may benefit English speakers more than others. It may amplify inequalities through access to compute and data.

A quote that fails the scope test is usually being used as a tribal token. The correction is to restate it with explicit conditions: “For high-stakes factual claims without retrieval, it behaves like a blurry JPEG.” “In adversarial settings, its stupidity can be exploited.” “For long-horizon governance, worst-case tail risks deserve attention.”

8.5 Stakes test: what policy or moral conclusion the quote is trying to smuggle in

Finally, ask the most important question: what is the quote trying to make you do?

Weird quotes almost always come with an implied action:

Misnomer quotes often imply: don’t grant authority; keep humans accountable; regulate as automation.
Monster quotes often imply: slow down; restrict; treat as catastrophic risk; centralize oversight.
Tool quotes often imply: adopt with guardrails; integrate; train users; keep humans in the loop.
Deflation quotes often imply: relax; don’t panic; treat as normal software; avoid special regulation.
Mimicry quotes often imply: distrust; focus on data ethics; limit use in meaning-heavy domains.
Magic quotes often imply: demand interpretability; avoid high-stakes deployment; require rigorous evaluation.

The stakes test exposes whether the quote is serving as a Trojan horse. Sometimes the implied action is wise. Sometimes it is opportunistic. For example, alarmist metaphors can be used to justify power consolidation or censorship theater. Deflation metaphors can be used to justify reckless deployment without accountability. “It’s just X” can become an excuse to avoid governance. “It’s a demon” can become an excuse to avoid practical policy and retreat into narrative.

Once you see the implied action, you can separate the descriptive and prescriptive layers. Even if you agree with the moral conclusion, you can demand better mechanisms and better scope. Even if you disagree with the moral conclusion, you can still learn from the failure modes the quote highlights.

The overall discipline is simple: do not treat a weird quote as an endpoint. Treat it as the beginning of a structured inquiry. Context tells you what conversation it came from. Claim tells you what it asserts. Mechanism tells you what must be true for it to work. Scope tells you when it applies. Stakes tells you what it is trying to make you do. With those five tests, you can stay open without being gullible, skeptical without being cynical, and serious without being captured by slogans.

9. Building a “Weird Quotes About AI” Dataset

If weird quotes about AI are a cultural dataset already circulating through society, the next step is to treat them like a dataset on purpose. Not as a scrapbook, and not as a list of zingers, but as a structured corpus that can be analyzed for patterns: what metaphors dominate, which actors produce them, how they spread, and what they do to policy, trust, and institutional behavior. The value of building such a dataset is not academic novelty. It is epistemic hygiene. When a phrase becomes ubiquitous, it starts shaping decisions. A dataset lets us move from anecdote to structure.

A good dataset also exposes what is missing. If most quotable lines come from a narrow set of voices, the dataset will reveal that skew. If certain metaphor families correlate with certain incentives (marketing versus warning versus pedagogy), the dataset will show it. And if quotes mutate into slogans detached from their original intent, the dataset can track the mutation rather than merely lament it.

9.1 Collection strategy: interviews, essays, talks, papers, social media

The first design choice is scope. The dataset should prioritize quotes that meet two criteria: they are widely circulated or likely to be circulated, and they encode a reframing of AI that implies some broader claim about capability, meaning, risk, or governance. This is not a “best quotes” list. It is an influence-tracking corpus.

A practical collection strategy has multiple pipelines:

Interviews and podcasts. These are fertile sources because speakers are under time pressure and often produce compressed metaphors. The dataset should capture both the quote and enough surrounding transcript to reconstruct context: what question was asked, what qualifiers were included, and what examples were used. Where possible, store the time stamp or segment boundary so the quote can be revisited in its original form.

Essays, op-eds, and books. Longer-form writing often provides the “argument behind the quote.” These sources are essential because they reveal whether the phrase was meant as a teaching metaphor, a moral warning, or a technical critique. When a short phrase comes from a long essay, the dataset should store the passage-level context and the author’s stated intent.

Talks and panels. Conference talks, university lectures, and public debates are key because speakers often aim for memorable lines. These are also the sources most likely to be clipped and recirculated. The dataset should record the venue type (technical conference versus public forum), the audience type, and the presence of adversarial questioning.

Peer-reviewed papers and technical reports. Academic writing is less likely to produce memeable one-liners, but when it does, those lines often become especially authoritative in

public discourse. Even when the paper itself does not have a zinger, journalists and commentators may extract a sentence or paraphrase it into a quote-like summary. Capturing both the original phrasing and common paraphrases helps track how technical claims turn into slogans.

Social media. This is where quote laundering happens fastest. Social media collection should include: original posts; reposts with modified wording; images with overlaid text; and "quote cards" that strip attribution and context. Social media also supplies a distribution signal: how fast the quote spreads, who amplifies it, what communities adopt it, and which variants become dominant.

A minimal record for each item should include:

- The quote text as it appears in the source
- A normalized "canonical" form (if the quote appears in multiple variants)
- Speaker/author
- Source type (interview, essay, talk, paper, social post)
- Date and approximate time period
- Audience/venue (technical, policy, general public, investor, etc.)
- The question or trigger (if applicable)
- A context excerpt (a paragraph or several sentences around it)
- Notes on whether it is direct quote, paraphrase, or reconstruction

The dataset should also explicitly store uncertainty. Many viral quotes are attribution-ambiguous. Treat that ambiguity as data rather than as a nuisance: "attributed to X," "primary source found/not found," "earliest traceable instance."

9.2 Annotation scheme: metaphor family, intent, target, implied claim

A dataset becomes analytically useful only when each item is annotated in a consistent way. The goal is not to overformalize, but to make the implicit structure explicit.

At minimum, annotate each quote along four dimensions:

Metaphor family. Assign one or more categories from the taxonomy: misnomer, compression, mimicry, magic, monster, tool, deflation, inversion. Multiple labels are often appropriate because quotes blend frames. For example, "AI is just statistics in a tuxedo" is both deflation and misnomer, with a hint of inversion.

Intent. What is the speaker trying to do? Suggested intent tags include: demystify, warn, criticize hype, criticize ethics, build a brand, teach, defend a paradigm, motivate regulation, discourage adoption, encourage adoption-with-guardrails. Intent is not always explicit, but context often reveals it.

Target. What is the quote aimed at? Possibilities: public misunderstanding, marketing language, a rival research approach, policymakers, overconfident users, institutional adoption, or the technology itself. “Stochastic parrot” targets epistemic trust and social risk. “Alchemy” targets research rigor and interpretability. “Demon” targets complacency about catastrophic risk.

Implied claim. Translate the quote into one or more plain-language propositions. This is the key step that makes the dataset evaluable. For example:

- “Blurry JPEG” implies: outputs are reconstructed from patterns; fine detail is unreliable; hallucination is expected without grounding.
- “Autocomplete on steroids” implies: core behavior is continuation; apparent reasoning can be an artifact of scale and context.
- “Neither artificial nor intelligent” implies: AI is materially and socially embedded; labor and extraction are central; the label hides responsibility.

Optionally, add two more annotations that often matter:

Normative payload. What action does the quote push toward? “Regulate,” “slow down,” “don’t anthropomorphize,” “treat as tool,” “don’t trust,” “adopt with verification.”

Confidence or evidence posture. Is the quote presented as a strong assertion, a tentative metaphor, a rhetorical flourish, or a deliberate provocation? This helps separate pedagogical scaffolding from ideological slogans.

The dataset should allow inter-annotator disagreement. Disagreement is not failure; it reveals ambiguity in the quote’s function. Over time, annotation debates can become a meta-dataset about how different communities interpret the same lines.

9.3 Measuring drift: how quotes mutate over time

Quote drift is not an accident; it is a predictable evolutionary process. Phrases mutate to maximize shareability, emotional punch, and ideological utility. A dataset can measure drift by tracking variant lineages.

There are several drift modes worth recording:

Wording simplification. Nuance gets shaved off. Qualifiers disappear. “In many contexts” becomes “always.” “May be” becomes “is.” Technical terms get replaced by metaphors. The quote becomes easier to repeat and harder to contest.

Target expansion. A quote aimed at a particular system becomes a general statement about all AI. “LLMs are like X” becomes “AI is X.” This expansion increases rhetorical power but decreases accuracy.

Attribution drift. Names attach to quotes because prestige increases transmission. A line said by a lesser-known speaker gets re-attributed to a famous figure. Or a paraphrase becomes a “direct quote.” The dataset should log earliest known attribution, later attributions, and the spread of each.

Moralization drift. Descriptive metaphors acquire moral force. “Blurry JPEG” becomes “deceptive.” “Parrot” becomes “fraud.” “Autocomplete” becomes “worthless.” These shifts matter because they change institutional responses.

Platform formatting drift. Quotes become image macros, short captions, or clipped video segments. Formatting affects meaning. An image quote card often strips the surrounding explanation that makes the metaphor responsible.

To measure drift, you can track:

- Edit distance or semantic similarity between variants
- Frequency over time of each variant
- Which variant becomes dominant in which community
- Co-occurrence with policy keywords (regulation, ban, safety, jobs)
- Co-occurrence with emotional language (fear, hype, fraud, doom)

This turns quote analysis into an empirical study of rhetoric as an information contagion. The payoff is practical: you can predict which metaphors are shaping governance discourse and which are fading.

9.4 Limitations: selection bias, survivorship bias, and translation issues

No quote dataset is neutral. The dataset is shaped by the same forces it aims to analyze.

Selection bias. If you collect from English-language media, you will overrepresent Anglophone frames and miss metaphors in other cultures. If you collect only from elite venues, you will miss

grassroots framings that may have more real impact. If you collect only viral quotes, you will miss careful quotes that shape expert discourse quietly.

Survivorship bias. The quotes that survive are those that travel. That means they are optimized for memorability, not for accuracy. Your dataset will therefore overrepresent extreme or vivid framings. This is not fatal, but it must be explicit: the dataset measures influence, not truth.

Attribution uncertainty. Many quotes are paraphrases, reconstructions, or composites. A dataset that pretends otherwise becomes a laundering engine itself. The dataset should include provenance flags: direct transcript, verified writing, secondary report, unsourced meme.

Translation issues. Metaphors do not translate cleanly. A parrot metaphor in one language may map onto a different animal or a different cultural association in another. A phrase that sounds deflationary in English may sound insulting or playful elsewhere. If the dataset spans languages, it must store original-language text and translator notes, not just English renderings.

Temporal instability. AI capabilities and deployment patterns evolve quickly. A quote that was fair in one era can become misleading later. The dataset should treat date as essential metadata, and analysis should avoid collapsing all time periods into a single "AI discourse" snapshot.

Interpretive ambiguity. Even with context, intent is not always clear. People use metaphors strategically; they may mean multiple things at once. Annotation should allow uncertainty and disagreement rather than forcing false precision.

Finally, there is the reflexive problem: building the dataset can itself amplify the quotes. A "weird quotes" paper can become a quote generator. That is not avoidable, but it can be managed by emphasizing reconstruction, context, and evaluation over anthology-style celebration.

A well-built dataset does not settle debates. It makes the debates legible. It shows which metaphors dominate, who benefits from them, how they mutate, and where the public conversation is being guided by slogans rather than by mechanisms, evidence, and accountable governance.

10. If "AI" Is a Misnomer, What Should We Call It Instead

The label "Artificial Intelligence" is a historical accident that became a global brand. It is also a category error generator. It invites people to assume mind where there is mechanism, agency where there is optimization, wisdom where there is pattern competence. That is why misnomer

quotes keep appearing: they are attempts to repair language so that expectations match reality and accountability has somewhere to land.

But naming is not just about philosophical correctness. Names are tools. They must be usable in law, procurement, journalism, education, and everyday conversation. If the replacement term is too technical, it will not travel. If it is too broad, it will reproduce the same confusion. If it is too deflationary, it will encourage complacency. If it is too dramatic, it will encourage panic. We are not choosing an abstract truth; we are choosing a social interface for a technology that is already reshaping institutions.

10.1 Candidate names and why each helps or fails

Below are common candidate labels—some already used, some proposed—along with what each clarifies and what each distorts.

Artificial cleverness. This is the most direct misnomer replacement. It preserves the sense that systems can be impressive without implying deep understanding. It is useful as a corrective to anthropomorphism. It fails as a broad label because "cleverness" sounds like party tricks, which understates the scale effects that matter for institutions. It also does not naturally cover non-language systems such as vision models, robotics, or scientific optimizers.

Statistical learning systems. This emphasizes mechanism and lineage. It helps remind readers that the core is learning from data. It fails in everyday speech because it is technical and "cold," and it can sound like a narrow subfield rather than a general class of systems. It also subtly implies that the systems are only about statistics, which can become a rhetorical minimization rather than a useful description.

Machine learning. This is the existing alternative most people already accept. It is better than "AI" in that it highlights training and adaptation rather than mind. It still misleads because "learning" evokes a human-like process of grasping meaning. It also does not distinguish among very different systems—supervised models, reinforcement learners, foundation models, etc.—and it can be too broad to guide governance.

Large language model or foundation model. These are more specific, and specificity is often what policy needs. They help avoid the "AI as everything" problem. They fail when the phenomenon extends beyond language or beyond "foundation" style models. They also age quickly: as architectures diversify, "LLM" may become one subclass rather than the center.

Predictive text engines or generative prediction systems. These are useful for demystifying chat-style systems. They highlight the generative and probabilistic nature of outputs. They fail because they underdescribe the ways these systems are now coupled to tools, retrieval, and multi-step workflows. They also sound dismissive even when used neutrally, which can polarize.

Automated decision systems. This label is governance-oriented. It focuses on what matters in high-stakes contexts: decisions, recommendations, triage, and ranking. It helps regulators and institutions target oversight to harm pathways. It fails because not all AI is used for decision-making; some is creative assistance, search, compression, or simulation. It also risks normalizing the idea that automated systems should be making decisions at all.

Algorithmic systems or computational models. These are extremely broad and therefore safe but not useful. They do not distinguish what is new about modern AI. They can collapse the conversation into "everything is algorithms," which evades the specific issues of scale, data, and emergent behaviors.

Synthetic media systems. This highlights one major social risk: generated text, images, audio, and video. It is useful for misinformation policy. It fails because it treats AI as media production only, ignoring decision support, prediction, optimization, and scientific applications.

Cognitive automation. This phrase is often closer to the institutional reality. It highlights that the technology automates portions of cognitive labor: drafting, summarizing, classification, pattern detection, code generation, and workflow triage. It helps keep accountability human. It fails because "automation" can obscure the interactive nature of these systems and can make them sound more deterministic than they are.

Decision support systems. This is a pragmatic term for many deployments. It implies the system supports, not replaces, human judgment. It helps establish governance posture: verify, supervise, retain responsibility. It fails when systems are deployed as autonomous actors or when organizations pretend "support" while effectively letting the system decide.

Model-mediated services. This phrase foregrounds the product reality: people interact with a service shaped by a model plus policies, retrieval, tooling, and incentives. It is good for accountability because it refuses to treat "the model" as the whole system. It fails for everyday use because it is bureaucratic and abstract.

Adjunct intelligence or intelligence augmentation. These are tool metaphors. They are attractive for framing AI as a capability amplifier rather than a replacement mind. They fail when they smuggle in "intelligence" again, reintroducing the prestige and agency confusion. They also tend to oversell benefits and understate risks.

Each candidate term solves a piece of the problem and creates a new one. That suggests the solution is not a single replacement word, but a layered naming strategy: different names for different contexts, with explicit scopes.

10.2 The tradeoff: precision vs usability

Naming sits on a spectrum between precision and usability.

High precision tends to be technical and narrow. “Transformer-based foundation model trained with next-token prediction and reinforcement learning” is precise, but no one will use it in daily speech or policy debate. Even “large language model” is already too technical for some audiences, and it fails to cover non-language systems.

High usability tends to be broad and metaphorical. “AI” is maximally usable: two letters that can mean anything. That is why it won. But this usability comes at the cost of confusion. Broad labels allow rhetorical manipulation. They let marketers imply agency and critics imply fraud, both without specifying mechanisms or contexts.

Institutions need an intermediate band: terms that are specific enough to govern and general enough to be adopted. The best naming strategy is therefore not an attempt to find the perfect word. It is a decision about which level of granularity to require in which setting.

In everyday conversation, people can say “AI” and mean “that chat system.” In law and procurement, that is unacceptable. A policy definition must specify:

- what the system does (generate, classify, rank, recommend, decide)
- what data it relies on (training sources, retrieval sources)
- what autonomy it has (advisory vs action-taking)
- what domain it operates in (health, finance, education, employment, security)
- what oversight is required

A precise policy definition does not need a perfect name. It needs a schema.

10.3 A pragmatic naming proposal for policy and for everyday speech

A workable compromise is a two-layer naming regime: one for everyday speech and one for governance. The key is to prevent the everyday label from becoming a legal category.

Everyday speech proposal. Keep “AI” as the casual umbrella term, but normalize a habit of adding a second word that signals function:

- “AI text generator”
- “AI code assistant”
- “AI image generator”

- “AI recommender”
- “AI screening tool”
- “AI decision support”

This small habit reduces mystique. It pulls the conversation from “what is it” toward “what does it do,” which is where practical evaluation begins.

Policy and institutional proposal. Avoid “AI” as the governing noun. Use a functional taxonomy that treats “AI” as a colloquial label and governs “model-mediated automated functions.” In policy documents and procurement, require systems to be categorized by function and autonomy:

- Generative systems (produce text, images, audio, video)
- Predictive systems (forecast outcomes, detect anomalies)
- Ranking and recommendation systems (prioritize content, people, resources)
- Decision-support systems (offer suggestions that humans adopt or reject)
- Automated decision systems (directly trigger actions affecting rights, access, or safety)

Then add two mandatory modifiers:

- Domain modifier (employment, education, health, finance, housing, policing, military, etc.)
- Autonomy modifier (advisory, human-in-the-loop, human-on-the-loop, autonomous)

Under this proposal, the institution never procures “AI.” It procures “a model-mediated decision-support system for claims triage” or “a generative system for drafting customer responses with mandatory human review.” The language forces governance posture into the contract and prevents vendors from hiding behind the mystique of “AI” while quietly shipping a brittle pipeline.

The social payoff is subtle but real. When we stop naming the technology as a mind and start naming it as a function, we make room for the right questions: What is it allowed to do? Who is responsible? What evidence supports reliability? What audits exist? What harms are foreseeable? And what is the exit plan if it fails?

In that sense, the best replacement for “AI” is not a single new word. It is a discipline of naming that keeps scope, function, and accountability visible—so that clever systems do not accidentally inherit the authority reserved for wise ones.

11. Conclusion: Weird Quotes as a Mirror of Our Moment

Weird quotes about AI are not the fringe of the conversation. They are the conversation's pressure valves. When a technology moves faster than shared understanding, society reaches for portable language. When institutions feel the ground shifting under their authority, they reach for metaphors that stabilize or destabilize. When incentives reward hype and punish nuance, people answer with punchlines, inversions, and misnomers—because that is what survives. In that sense, the weirdness is not an embarrassment. It is evidence. It shows where the collective mind is straining to reconcile new capability with old categories, and where power is being contested through naming.

If you treat these quotes as data, a striking pattern emerges: the same metaphors recur because the same anxieties recur. Are we being manipulated by fluency? Are we granting authority without accountability? Are we drifting into dependence on systems we do not understand? Are we facing catastrophic risk or merely a new wave of automation? The quotes cluster around those questions because those questions are the true substrate of the debate.

11.1 What the quote ecology reveals about trust and power

The ecology of weird quotes reveals that trust is the central resource being fought over. AI systems do not merely perform tasks; they ask to be believed. And belief—epistemic trust, moral trust, institutional trust—is a scarce and powerful commodity.

One cluster of quotes is essentially an anti-trust campaign: "stochastic parrots," "blurry JPEG," "autocomplete," "curve fitting," "alchemy," "stupidity." These are counter-spells against unjustified confidence. They target a specific risk: that the public and institutions will treat fluency as knowledge and coherence as truth. In these quotes, the system is not evil; it is unreliable in ways that become dangerous when paired with authority and scale.

Another cluster of quotes is an anti-accountability campaign, though often unintentionally. "Black box," "emergent," "we can't explain," and certain forms of mystique can function as responsibility fog. When a system is framed as mysterious, it becomes easier for organizations to claim that failures are unavoidable. This is not always the speaker's intention—often it is a valid statement of epistemic limits—but institutions can weaponize the mystique as cover.

A third cluster is a power-warning campaign: "neither artificial nor intelligent" and related materialist framings. These quotes insist that the real story is not mind, but infrastructure: energy, compute concentration, labor, and governance. They reveal a fear not only of what models can do, but of who controls them, who profits, and who becomes dependent.

Finally, the monster metaphors reveal a trust breakdown in the future itself. They are expressions of a belief that the normal institutions of risk management may not scale to the

pace and stakes of AI. They imply that the public cannot trust incrementalism, and that conventional governance will arrive too late.

Put bluntly: the quote ecology is a proxy war over legitimacy. Who gets to define reality? Who gets to distribute responsibility? Who gets to set the rules? Weird quotes are the cultural signatures of that war.

11.2 The durable lesson: replace slogans with operational questions

The most durable lesson is methodological. Weird quotes are not answers; they are prompts. Their proper use is to generate operational questions that can be tested, governed, and audited. If you can translate a quote into a checklist, you have converted rhetoric into responsibility.

Replace “It’s just autocomplete” with:

- What is the system’s error profile in our domain?
- How will humans verify outputs, and what is the fallback when verification fails?
- What training do users receive to prevent overtrust?
- What security controls exist against prompt injection and data exfiltration?

Replace “Blurry JPEG of the web” with:

- Is the system grounded in retrieval for factual claims?
- Does it cite sources, and are citations verifiable?
- What kinds of prompts trigger hallucination, and what guardrails reduce it?
- What are the acceptable error rates for our use-case?

Replace “Stochastic parrots” with:

- Where does the system produce persuasive but wrong content?
- How do we detect bias amplification and stereotyping?
- What human review is required in meaning-heavy domains?
- What disclosure and transparency is owed to users?

Replace “Alchemy” with:

- What evidence exists for reliability under distribution shift?
- What is the red-team record and incident history?

- What monitoring exists in production, and who owns response?
- What interpretability or evaluation artifacts are required for deployment?

Replace “Summoning the demon” with:

- What is the credible threat model at current capability levels?
- Which deployment classes create catastrophic tail risks?
- What is the governance mechanism for capability jumps?
- Who has authority to pause, restrict, or roll back deployment?

Replace “AI is neither artificial nor intelligent” with:

- What are the data provenance and labor conditions behind the system?
- What energy and environmental costs are externalized?
- What concentration risks exist in the supply chain and compute access?
- What accountability chain exists from designer to deployer?

This translation discipline is the antidote to both credulity and cynicism. It refuses to let slogans do the thinking, but it also refuses to dismiss slogans as meaningless. It treats them as signals that something important is being compressed—and therefore something important must be made explicit.

11.3 An agenda for clearer speech in the centaur era

The “centaur era” is the era in which human cognition is routinely coupled to machine-generated suggestions, summaries, plans, and drafts. In such an era, language becomes a form of infrastructure. How we speak about AI will partly determine how we build with it and how we are governed by it.

A practical agenda for clearer speech has several components:

First, normalize functional naming. In everyday talk, keep “AI” if necessary, but add the function: generator, recommender, classifier, decision-support. In policy and procurement, prohibit “AI” as a sufficient description. Require functional category plus domain plus autonomy level. Language should force accountability.

Second, institutionalize the separation of descriptive and prescriptive claims. When a speaker says “It doesn’t understand,” ask whether they mean “it is unreliable in truth-tracking,” “it lacks grounding,” or “it lacks consciousness.” When a speaker says “It’s dangerous,” ask whether they mean “high error cost,” “misuse at scale,” or “catastrophic tail risk.” This separation prevents

philosophical debates from hijacking practical governance and prevents practical failures from being dismissed as “just philosophy.”

Third, treat metaphors as partial lenses, not total theories. A healthy discourse explicitly asks: what does this metaphor illuminate, and what does it hide? When a metaphor is offered, it should come with its blind spot. That simple habit turns metaphor from ideology into pedagogy.

Fourth, build public literacy around the agency illusion. Conversational interfaces make machines feel like minds. That feeling is not a sin; it is a human reflex. But institutions should train users to distinguish politeness from truth, confidence from calibration, and fluency from authority. The core skill is not skepticism as attitude, but verification as habit.

Fifth, demand accountability for the whole system, not just the model. Most harms arise in the system-of-use: incentives, UI design, data pipelines, monitoring, and human workflows. Clear speech should keep that system visible. The phrase “the AI did it” should be treated as a red flag, not a conclusion.

If this agenda sounds modest, that is the point. The future of AI will be shaped not only by breakthroughs in architecture and compute, but by the social technologies we build around them: language, norms, institutions, and governance. Weird quotes will continue to appear because they are how culture metabolizes novelty. The goal is not to end the weirdness. The goal is to turn it into clarity—so that clever machines do not inherit moral authority by accident, and so that humans do not surrender responsibility under the spell of their own metaphors.

12. References

12.1 Primary sources: original talks, interviews, essays, transcripts

- Penrose, Roger. “Why I Never Quite Liked the Word ‘Artificial Intelligence’ … ‘Artificial Cleverness’ Would Be OK.” AI for Good Global Summit (ITU), transcript (May 15, 2018). https://aiforgood.itu.int/wp-content/uploads/2018/05/Roger_Penrose_-_AI_for_Good_Global_Summit_Transcript.pdf
- Oxford Union. Clip reposting Penrose on “Artificial intelligence is a misnomer” (video excerpt; attribution aid). https://x.com/OUAiSociety/status/1724664992275153140
- Penrose, Roger. The Emperor’s New Mind. Oxford University Press, 1989.
- Penrose, Roger. Shadows of the Mind. Oxford University Press, 1994.
- Chiang, Ted. “ChatGPT Is a Blurry JPEG of the Web.” The New Yorker (Feb 9, 2023). https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web
- Bender, Emily M.; Gebru, Timnit; McMillan-Major, Angelina; Shmitchell, Margaret. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” FAccT ’21 (2021). https://doi.org/10.1145/3442188.3445922
- Crawford, Kate. “Microsoft’s Kate Crawford: ‘AI is neither artificial nor intelligent’.” The Guardian (Jun 7, 2021). https://www.theguardian.com/technology/2021/jun/06/microsofts-kate-crawford-ai-is-neither-artificial-nor-intelligent
- Crawford, Kate. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, 2021.
- Rahimi, Ali. “Machine Learning Has Become Alchemy.” NeurIPS Test-of-Time Award talk (Dec 2017). https://www.youtube.com/watch?v=x7psGHgatGM
- Hutson, Matthew. “AI researchers allege that machine learning is alchemy.” Science (May 3, 2018). https://www.science.org/content/article/ai-researchers-allege-machine-learning-alchemy
- Pearl, Judea. Interview excerpt emphasizing “curve fitting” vs causal models (source anthology). https://www.phenomenalworld.org/sources/even-with-closed-eyes/
- Dijkstra, Edsger W. “On IPW’s” (EWD 867) (includes “submarines can swim” line). EWD Transcriptions (UT Austin). https://www.cs.utexas.edu/~EWD/transcriptions/EWD08xx/EWD867.html

- Dijkstra, Edsger W. “The threats to computing science” (EWD 898) (related context). EWD Transcriptions (UT Austin). https://www.cs.utexas.edu/~EWD/transcriptions/EWD08xx/EWD898.html
- McCarthy, John; Minsky, Marvin; Rochester, Nathan; Shannon, Claude. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence” (1955). https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html
- Turing, Alan M. “Computing Machinery and Intelligence.” Mind 59(236), 1950.
- Musk, Elon. “With artificial intelligence we are summoning the demon.” (coverage of remarks; attribution aid). https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/
- Hawking, Stephen. “AI could spell end of the human race” (BBC interview coverage; attribution aid). https://www.theguardian.com/science/2014/dec/02/stephen-hawking-intel-communication-system-astrophysicist-software-predictive-text-type
- Clarke, Arthur C. Profiles of the Future. Victor Gollancz, 1962.
- Hofstadter, Douglas. Gödel, Escher, Bach. Basic Books, 1979. (often cited in “AI effect” discussions)
- Quote Investigator. “As Soon As It Works, No One Calls It AI Anymore.” (Tesler’s Theorem provenance; attribution aid). https://quoteinvestigator.com/2024/06/20/not-ai/
- Weizenbaum, Joseph. Computer Power and Human Reason: From Judgment to Calculation. W.H. Freeman, 1976.
- Dreyfus, Hubert L. What Computers Can’t Do. Harper & Row, 1972.

12.2 Technical background: ML fundamentals, LLM behavior, evaluation science

- Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron. Deep Learning. MIT Press, 2016. https://www.deeplearningbook.org/
- Bishop, Christopher M. Pattern Recognition and Machine Learning. Springer, 2006.
- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. The Elements of Statistical Learning (2nd ed.). Springer, 2009.
- Sutton, Richard S.; Barto, Andrew G. Reinforcement Learning: An Introduction (2nd ed.). MIT Press, 2018. http://incompleteideas.net/book/the-book-2nd.html

- Sutton, Richard S. “The Bitter Lesson.” (essay; 2019). http://www.incompleteideas.net/IncIdeas/BitterLesson.html
- Vaswani, Ashish, et al. “Attention Is All You Need.” NeurIPS 2017. https://arxiv.org/abs/1706.03762
- Brown, Tom B., et al. “Language Models are Few-Shot Learners.” NeurIPS 2020. https://arxiv.org/abs/2005.14165
- Ouyang, Long, et al. “Training language models to follow instructions with human feedback.” 2022. https://arxiv.org/abs/2203.02155
- Wei, Jason, et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” 2022. https://arxiv.org/abs/2201.11903
- Kojima, Takeshi, et al. “Large Language Models are Zero-Shot Reasoners.” 2022. https://arxiv.org/abs/2205.11916
- Wang, Xuezhi, et al. “Self-Consistency Improves Chain of Thought Reasoning in Language Models.” 2022. https://arxiv.org/abs/2203.11171
- Lewis, Patrick, et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” NeurIPS 2020. https://arxiv.org/abs/2005.11401
- Yao, Shunyu, et al. “ReAct: Synergizing Reasoning and Acting in Language Models.” 2022. https://arxiv.org/abs/2210.03629
- Schick, Timo, et al. “Toolformer: Language Models Can Teach Themselves to Use Tools.” 2023. https://arxiv.org/abs/2302.04761
- Liang, Percy, et al. “Holistic Evaluation of Language Models (HELM).” 2022. https://arxiv.org/abs/2211.09110
- Srivastava, Aarohi, et al. “BIG-bench: Beyond the Imitation Game benchmarking.” 2022. https://arxiv.org/abs/2206.04615
- Hendrycks, Dan, et al. “Measuring Massive Multitask Language Understanding (MMLU).” 2021. https://arxiv.org/abs/2009.03300
- Lin, Stephanie, et al. “TruthfulQA: Measuring How Models Mimic Human Falsehoods.” 2021. https://arxiv.org/abs/2109.07958
- OpenAI. “GPT-4 Technical Report.” 2023. https://arxiv.org/abs/2303.08774
- NIST. “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations.” NISTIR 8269 (2020). https://doi.org/10.6028/NIST.IR.8269

- Mitchell, Margaret, et al. “Model Cards for Model Reporting.” FAT* ’19 (2019). https://doi.org/10.1145/3287560.3287596
- Gebru, Timnit, et al. “Datasheets for Datasets.” CACM 64(12), 2021. https://doi.org/10.1145/3458723

12.3 Philosophy of mind and cognition: understanding, meaning, intentionality

- Searle, John R. “Minds, Brains, and Programs.” Behavioral and Brain Sciences 3(3), 1980.
- Harnad, Stevan. “The Symbol Grounding Problem.” Physica D 42, 1990.
- Dennett, Daniel C. Consciousness Explained. Little, Brown, 1991.
- Dennett, Daniel C. The Intentional Stance. MIT Press, 1987.
- Putnam, Hilary. “Brains and Behavior.” (classic functionalism discussions; various essays).
- Fodor, Jerry. The Language of Thought. Harvard University Press, 1975.
- Chalmers, David J. The Conscious Mind. Oxford University Press, 1996.
- Clark, Andy; Chalmers, David. “The Extended Mind.” Analysis 58(1), 1998.
- Wittgenstein, Ludwig. Philosophical Investigations. 1953.
- Davidson, Donald. “Mental Events.” (anomalous monism; 1970).
- Dretske, Fred. Knowledge and the Flow of Information. MIT Press, 1981.
- Brandom, Robert. Making It Explicit. Harvard University Press, 1994.
- Floridi, Luciano. The Philosophy of Information. Oxford University Press, 2011.
- Haugeland, John. Artificial Intelligence: The Very Idea. MIT Press, 1985.

12.4 Media studies and rhetoric: metaphor, framing, memetics, quote drift

- Lakoff, George; Johnson, Mark. Metaphors We Live By. University of Chicago Press, 1980.
- Entman, Robert M. “Framing: Toward Clarification of a Fractured Paradigm.” Journal of Communication 43(4), 1993.
- Tversky, Amos; Kahneman, Daniel. “The Framing of Decisions and the Psychology of Choice.” Science 211(4481), 1981.
- McCombs, Maxwell E.; Shaw, Donald L. “The Agenda-Setting Function of Mass Media.” Public Opinion Quarterly 36(2), 1972.

- Lippmann, Walter. Public Opinion. 1922.
- Dawkins, Richard. The Selfish Gene. Oxford University Press, 1976. (memes concept origin)
- Shifman, Limor. Memes in Digital Culture. MIT Press, 2014.
- Sunstein, Cass R. On Rumors. Farrar, Straus and Giroux, 2009.
- Allport, Gordon W.; Postman, Leo. The Psychology of Rumor. Henry Holt, 1947.
- Boyd, danah. "Social Network Sites as Networked Publics." (in Papacharissi, Networked Self; 2010).
- Tufekci, Zeynep. Twitter and Tear Gas. Yale University Press, 2017.
- O'Toole, Garson. Quote Investigator (methodological archive on misattribution/quote mutation). https://quoteinvestigator.com/
- Phillips, Whitney; Milner, Ryan M. The Ambivalent Internet. Polity, 2017.
- Marwick, Alice; Lewis, Rebecca. "Media Manipulation and Disinformation Online." Data & Society, 2017. https://datasociety.net/library/media-manipulation-and-disinfo-online/

12.5 Governance and ethics: definitions, accountability, auditability, procurement

- NIST. AI Risk Management Framework 1.0 (AI RMF 1.0). NIST AI 100-1 (Jan 2023). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
- White House (OSTP). Blueprint for an AI Bill of Rights (Oct 2022). https://www.govinfo.gov/app/details/GOVPUB-PREX23-PURL-gpo193638
- U.S. Executive Office. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of AI (Oct 30, 2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
- OMB. Memorandum M-24-10: Advancing Governance, Innovation, and Risk Management for Agency Use of AI (Mar 28, 2024). https://www.whitehouse.gov/omb/memoranda/m-24-10/
- OECD. Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449) (May 22, 2019). https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

- European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Official Journal text (June 13, 2024 / OJ publication July 2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
- European Commission. “AI Act enters into force” (Aug 1, 2024). https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en
- ISO/IEC. ISO/IEC 23894:2023 Artificial intelligence — Guidance on risk management. (standard landing page). https://www.iso.org/standard/77304.html
- ISO/IEC. ISO/IEC 42001:2023 Artificial intelligence management system (AIMS). https://www.iso.org/standard/81230.html
- IEEE. IEEE 7000 series (ethical considerations / transparency / bias standards family). https://standards.ieee.org/industry-connections/ec/7000/
- EU High-Level Expert Group on AI. Ethics Guidelines for Trustworthy AI (2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- Diakopoulos, Nicholas. Automating the News: How Algorithms Are Rewriting the Media. Harvard University Press, 2019.
- Pasquale, Frank. The Black Box Society. Harvard University Press, 2015.
- Raji, Inioluwa Deborah, et al. “Closing the AI Accountability Gap.” FAT* ’20 (2020). https://doi.org/10.1145/3351095.3372873
- Selbst, Andrew D., et al. “Fairness and Abstraction in Sociotechnical Systems.” FAT* ’19 (2019). https://doi.org/10.1145/3287560.3287598

The author has published over 4,900 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.


"Stochastic Parrot"
"Blurry JPEG of the Web"
Artificial Cleverness"
"Summoning the Demon?"
AI EFFECT